Enhancing Sensor Performance with Statistical Data Analytics

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

James Scrimgeour Wright

July 2021

Abstract

This thesis examines the use of Automatic Identification System (AIS) information to generate a picture of maritime activity. It derives suitable methods to produce tracks of vessel movements, both in littoral and open-ocean scenarios, removing ambiguities and highlighting doppelgänger. The thesis then goes on to describe techniques to improve our understanding of maritime activities through the extraction of individual vessel behaviours and the generation of models describing normal behaviours to highlight abnormalities.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Simon Maskell and my co-supervisor Stephen Ablett for all their support and invaluable guidance throughout the PhD and this thesis.

I would like to thank many. I would like to thank my friends; Ross, Jane, those from the Research group; particularly, Al, Chinmay and Chloë; and those from the music circles; Tom, John, and the members of the sax groups, wind bands and big bands I have used as an evening's escape from the thesis; the team and friends in the University's Walking for Health group; and, of course, not forgetting the inexplicable members of the *League of Evil*.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

A portion of the research in chapter 5 is published as;

Zhou, Y., Wright, J. and Maskell, S. (2019) *A Generic Anomaly Detection Approach Applied to Mixture-of-Unigrams and Maritime Surveillance Data*, Sensor Data Fusion Symposium 2019. `https://doi.org/10.1109/SDF.2019.8916633` [148]

A portion of the research covered in Chapters 2 and 3 was presented as a poster at the Defence and Security Doctoral Symposium 2018;

James Wright (2018) *Disambiguating cooperative sensor data.* Poster. Defence and Security Doctoral Symposium 2018. Cranfield University (2018): 2018 Defence and Security Doctoral Symposium (DSDS18) in conjunction with Dstl and AWE: Symposium outputs. figshare. Collection. `https://doi.org/10.17862/cranfield.rd.c.4177190.v8` [144]

Contents

Abstract i

Acknowledgements ii

Declaration iii

Contents vii

List of Figures xiii

List of Tables xiv

1 Introduction 1
- 1.1 Thesis Overview . . . 2
- 1.2 Novel Contributions of this Thesis . . . 3
- 1.3 Publications . . . 4

2 Background 5
- 2.1 State Estimation and Tracking . . . 5
  - 2.1.1 Bayesian statistics . . . 5
  - 2.1.2 State Estimation . . . 6
    - 2.1.2.1 Transition Equation . . . 6
    - 2.1.2.2 Measurement Equation . . . 7
    - 2.1.2.3 The Kalman Filter . . . 8
    - 2.1.2.4 Motion Models . . . 9
  - 2.1.3 Tracking . . . 12
    - 2.1.3.1 Tracking individual targets . . . 17
    - 2.1.3.2 Measurement Space . . . 17
    - 2.1.3.3 Target State Space . . . 18
    - 2.1.3.4 Measurement Equation . . . 19
    - 2.1.3.5 Measurement Noise . . . 20

2.2 Track Stitching . . . 20
2.2.1 Forward Prediction . . . 21
2.2.2 The Rauch-Tung-Striebel Smoother . . . 27
2.3 Text Analytics . . . 28
2.3.1 Adaptive Grid and Geospatial Clustering . . . 28
2.3.2 Introduction to Probabilistic Topic Models . . . 31
2.3.2.1 Mixture-of-Unigrams - *The Dirichlet Multinomial Mixture model* . . . 31
2.3.2.2 The Latent Dirichlet Allocation Model . . . 35
2.3.2.3 Specific instances of the approach for Latent Dirichlet Allocation . . . 36
2.4 Change Point Detection . . . 37
2.4.1 Change Point Detection from Simulated Count Data . . . 39
2.5 The Maritime Challenge . . . 41
2.6 Maritime Data and Sensor Types . . . 45
2.6.1 Non-Cooperative Data . . . 45
2.6.1.1 Radar . . . 45
2.6.1.2 Space-based Radar . . . 45
2.6.1.3 Radio Direction Finding . . . 46
2.6.2 Legislation . . . 46
2.6.2.1 Safety of life at Sea (SOLAS) . . . 47
2.6.2.2 International Maritime Organisation . . . 47
2.6.3 Databases . . . 47
2.6.3.1 IHS Markit World Registry of Shipping . . . 47
2.6.3.2 List of Ports/UN Locodes . . . 48
2.6.4 Cooperative Data . . . 51
2.6.4.1 Automatic Identification System . . . 51
2.6.4.2 Long Range Tracking and Identification . . . 51
2.6.4.3 Vessel Monitoring System . . . 52
2.6.4.4 Notice of Port Arrival . . . 52
2.7 Automatic Identification System . . . 52
2.7.1 Message Types . . . 53
2.7.2 Message Structure . . . 54
2.7.3 Payload Content . . . 54
2.7.4 The MMSI Number . . . 55
2.7.5 The IMO Registration Number . . . 55
2.7.6 AIS Ship Type . . . 56
2.8 Datasets used in this study . . . 57
2.8.1 Merseyside Dataset . . . 58
2.8.2 North Atlantic Dataset . . . 61
2.8.3 Global Dataset . . . 64

2.8.4 Peculiarities . . . 70
2.9 Analysis of Error Messages . . . 71
2.9.1 Analysing the Tracklets for Errors . . . 73
2.9.2 Discussion of Findings . . . 79
2.10 Summary . . . 79

3 Disambiguation **80**
3.1 Introduction . . . 80
3.2 Quantification of performance . . . 83
3.2.1 Normal behaviour of a single vessel reporting on a single MMSI . . . 83
3.3 Applying the Tracker to AIS data . . . 92
3.4 Results . . . 95
3.5 Discussion . . . 102

4 Single Ship Analysis **103**
4.1 Track Joining and Reflagging . . . 103
4.1.1 Results for Reflagging . . . 107
4.2 Ship Stopping . . . 108
4.2.1 Utilising a by-product of the multiple target tracker . . . 108
4.2.2 Ports arrivals and departures . . . 110
4.2.3 Discussion . . . 112

5 Multi-Ship Analysis **113**
5.1 Behavioural Detection . . . 113
5.1.1 Latent Dirichlet Allocation analysis of AIS data . . . 113
5.1.1.1 Mixture of Unigram Anomaly detection for AIS data using symbolic positions . . . 115
5.1.1.2 Mixture-of-Unigram Anomaly Detection from Maritime Surveillance Data . . . 117
5.2 Change Point Detection . . . 121
5.2.1 Assessment of performance . . . 121
5.2.2 Results . . . 126
5.2.3 Discussion . . . 132

6 Conclusions and Recommendations **133**
6.1 Summary of Thesis Contributions . . . 133
6.2 Recommendations for Future Work . . . 134
6.2.1 Extensions to Work within the Thesis . . . 134
6.2.2 New Directions Motivated by the Thesis . . . 135
6.2.2.1 Prioritising Targets on the Basis of the Passage of Time and Machine Learning . . . 135

6.2.2.1.1 Survival Analysis and Censored Data 139
6.2.2.2 Closing Remarks 143

A Collaborations **144**
A.1 Track Analytics 144
A.2 Collaboration with Dstl and the National Maritime Information Centre . . 145
A.3 Stone Soup 146
A.4 Project NELSON 146

B AIS Specifications **147**
B.1 Message Types 147
B.2 Message Variables 148
B.3 Navigation Status 152
B.4 Rate of Turn 152
B.5 Ship Type 153

C General Information **156**
C.1 List of Acronyms 156

D Gibbs sampling **159**
D.1 The Gibbs Sampler Algorithm 159

E Additional Information **161**
E.1 Taylor Series Expansion 161

References **164**

List of Figures

2.1 The Multi-Target Tracking Pipeline. The blue boxes represent the state estimation components discussed in Section 2.1.2 and the red boxes represent the track management and data association components associated with multiple target tracking discussed in Section 2.1.3. 13
2.2 An example where the entire vessel trajectory is a single tracklet compared to the vessel trajectory being split over multiple tracklets. 22
2.3 The results of overfitting from Figure 2.2 applied to a different MMSI. . . . 23
2.4 Fusion from sensor disambiguation to centralised reflagging. 24
2.5 The process of using the automatic tracklet join followed by the reflagging method on a set of disambiguated tracklets. 26
2.6 Adaptive grid applied to the UK subset of the global dataset. 29
2.7 An example of how to convert a latitude-longitude track into a sequence of symbols. In this track, "$32 : 2, 50 : 10, \ldots$" is a part of the extracted data (i.e., document) and means that the ship stayed in region 32 for two AIS messages and stayed in region 50 for 10 AIS messages. Note that as the interval between two consecutive AIS messages is one hour, two AIS messages from the same region imply that the ship stayed in the region for about two hours. 30
2.8 The graphical representation of the Mixture-of-Unigrams model. The circles mean samples drawn from a distribution. The boxes mean a number of replications. α and η are the parameters for Dirichlet Distributions from which the multinomial distributions are drawn. θ and β define multinomial distributions that represent the topic and word distributions, respectively. z and w are samples that are drawn from the corresponding multinomial distributions. 33
2.9 Graphical model representation of the Latent Dirichlet Allocation model . . 35
2.10 400 observations of Poisson change point data x with abruptly changing mean. The points represent observations drawn from the underlying generating mean, λ with values 2, 20 and 4 38

2.11 Simulation of change point data and the total number of changes detected by the algorithm for 100 tests. (Jitter applied to the integer data.) 40
2.12 Simulation of change point data with a single change point and two differing distributions and the log posterior of the data produced by the algorithm for 100 tests. (Jitter applied to the integer data.) 41
2.13 Visualisation of land and inland water (black), territorial waters [12nm] (white), exclusive economic zone [200nm] (dark blue), and international waters (light blue). 43
2.14 Distribution of vessels (unique MMSI numbers) per degree longitude (dark blue) against the percentage of ocean per longitude (light blue). 44
2.15 Distribution of vessels (unique MMSI numbers) per degree latitude (dark blue) against the percentage of ocean per latitude (light blue). 44
2.16 The ports defined in the United Nations Code for Trade and Transport Locations grouped by country. 49
2.17 Duplicate port names in the United Nations Code for Trade and Transport Locations Database. The depicted 371 ambiguous ports outline the discrepancy of using a text box to name a location over using a system such as the UN/LOCODE as a destination for the Destination field used in the static AIS messages. 50
2.18 Example of AIS Message Types 1 and 3 54
2.19 A 24 hour period from a single AIS receiver at Fort Perch Rock located at the Northernmost tip of the Wirral Peninsula. Density resolution at 2 arcseconds. *(Data provided by Denbridge Marine)* 58
2.20 The figure represents the maximum distance a vessel travels over a given time from the Merseyside dataset. The blue depicts the 95% confidence interval and the dark blue depicts the 68% interval, where 68% of all ships are in the dark blue and 95% are in the light blue derived from the vessel maximum speed in the WRS database. The red region depicts vessels travelling faster than the water speed record of (511kmph). 59
2.21 Elapsed time between subsequent AIS messages of the same MMSI. 60
2.22 Count of MMSIs with valid MID number per flag state for the Merseyside dataset. 60
2.23 A 32 day period from an aggregated commercial AIS data source for the North Atlantic. Density resolution at 3 arcminutes. *(Data provided by Exact Earth)* 61
2.24 Count of MMSIs with valid MID number per flag state for the 23 most common countries for the North Atlantic dataset. 62
2.25 Count of the number of IMOs per MMSI for the North Atlantic dataset. . . 62
2.26 Count of the number of MMSIs per IMO for the North Atlantic dataset. . . 63
2.27 Count of number of different lengths per MMSI reported in the static reports for the North Atlantic dataset. 63

2.28 Count of number of different beams per MMSI reported in the static reports for the North Atlantic dataset. 64
2.29 A 15 day period from an aggregated commercial AIS data source for the world where the AIS reports are hourly. Density resolution at 10 arcminutes. *(Data provided by IHS Markit)* 65
2.30 The IHS Dataset is has hourly reporting rates which means in the 15 day dataset a vessel can only have reported 360 times. The vertical red line denotes the split between those quantity of vessels with their total message count below this threshold and the quantity of vessels with more than 360 reports. 67
2.31 Frequency of distances between consecutive observations. 68
2.32 For a given MMSI, the count here is for the number of multiple IMOs. . . . 68
2.33 For a given IMO, the count here is for the number of multiple MMSIs. . . . 69
2.34 MMSI Exemplar 71
2.35 The count of percentage error per MMSI. This graph shows that for the majority of vessels position reports are received with no errors however there are more than 1000 vessels that report a Null position report. 74
2.36 The count of time interval between consecutive observations for the Merseyside dataset. The majority of observation update rates are as expected less than 5 minutes. 75
2.37 The count of time interval between consecutive observations for the North Atlantic dataset. 75
2.38 The count of time interval between consecutive observations for the global dataset. 76
2.39 Map showing Merseyside dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 1 hour. 76
2.40 Map showing the North Atlantic dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 5 days. 77
2.41 Map showing the global dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours. 77
2.42 Map showing the Mediterranean extracted from the global data. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours. 78
2.43 Map showing the South China Seas extracted from the global data. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours. 78

3.1 MMSI 353203000 with erroneous ($91°N$, $181°E$) messages 81

3.2 MMSI 353203000 with erroneous ($91°N$, $181°E$) messages with observations joined. . . . 82
3.3 Example simulation of a single target with measurement noise $q = 10$. . . . 84
3.4 Example simulation of a single target with measurement noise $q = 1000$. . . 85
3.5 Example simulation of a single target with measurement noise $q = 100000$. . 86
3.6 Example simulation of a True vessel (blue), 1 additional vessel (green), in this example, spawning from a point along the true vessel trajectory, and 4 additional objects (orange). . . . 88
3.7 Simulation results of SIAP Completeness Metric (on all tracks and all vessels) vs the percent of actual measurements generated from true vessel for n probability of detecting true vessel over other. . . . 90
3.8 Simulation results of SIAP Completeness Metric (on all tracks and only the true vessel) vs the percent of actual measurements generated from true vessel for n probability of detecting true vessel over other. . . . 91
3.9 MMSI 353203000 with erroneous ($91°N$, $181°E$) messages with observations tracked by a Kalman filter. The yellow lines denote the raw data and the purple line represents the tracked vessel and the blue points represent the discarded observations. . . . 94
3.10 Result of initiating a new Kalman filter on AIS positions further away than existing Kalman filters. . . . 95
3.11 Subset of data from a single AIS receiver where each MMSI point has been joined into a path. . . . 96
3.12 Subset of interesting MMSIs from the North Atlantic dataset. . . . 97
3.13 Subset of interesting MMSIs from the Global dataset. . . . 98
3.14 The result of applying the disambiguation process to the Fort Perch Rock dataset. . . . 99
3.15 The result of applying the disambiguation process to the North Atlantic dataset. . . . 100
3.16 The result of applying the disambiguation process to the Global dataset. . . . 101

4.1 The ROC curve depicts the difference between using the predicting forward only, and the combination of using both forward and backward predictions. 105
4.2 The ROC curve depicts the difference between the number of possible tracklet pairs and their assignment using the combination of forward and backward predictions to generate the joining cost. . . . 106
4.3 Simulation of a vessel moving (thin blue line) and stopping (thick blue line) and the associated speed (red) calculated from the velocity components of the track states. . . . 108

4.4 ROC Curve depicting the TPR against the FPR for 100 simulations of a vessel moving and stopping (e.g., Figure 4.3) at varying degrees of noisy measurements, where 0 probability of noise refers to the case where there are 0 measurements from a measurement model with high noise ($q = 10000$) and 1 being where the 100% of measurements are from a measurement model with high noise. 109
4.5 Visualising the nearest ports to the stopped locations in Table 4.3. 111

5.1 The results of applying the Latent Dirichlet Allocation approach for detecting 10 behaviours to a subset of the Global dataset. This is behaviour 2. 114
5.2 The results of applying the Latent Dirichlet Allocation approach for detecting 10 behaviours to a subset of the Global dataset. This is behaviour 8. 115
5.3 The quad-tree adaptive grid used for converting the latitude-longitude coordinates to symbolic representations. Data points within any red rectangles have indices, the rest of the regions are not considered as they did not appear in the training set. Each cell contains less than $Q = 3000$ data points. There are 1183 cells in the grid. 116
5.4 The positions of ships that were reported in the AIS messages. 119
5.5 Some examples of detected anomalies using models trained with each of a number of types of ship. 120
5.6 25 simulated regions with change points ordered by geographical region ID, where red denotes the ground truth and black denotes the vessel count. The regions contain time series that remains mostly constant, 2, 21, 24), with some change (3, 12, 15) and large changes (8 22, 14). 122
5.7 25 simulated regions as seen in Figure 5.6 with the corresponding posterior values plotted. 123
5.8 The regions presented in Figure 5.6 ordered by the posterior score. 124
5.9 The equivalent plot to Figure 5.6 where the ground truth is used to calculate the posterior score. 125
5.10 The change point method applied to the region with ID 1530 with aggregated MMSI count grouped per hour. 128
5.11 The change point method applied to the augmented region with ID 1530 with aggregated MMSI count grouped per hour. Here the augmentation of zeros has been manually added. 129
5.12 The change point method applied to the region with ID 1611 with aggregated MMSI count grouped per hour. 130
5.13 The change point method applied to the augmented region with ID 1611 with aggregated MMSI count grouped per hour. Here the augmentation of zeros has been manually added. 131

6.1 Visualisation of time since last looked. 141

List of Tables

2.1 Summary of tracking filters and the cases where they are best applied [125] 16
2.2 AIS message type class breakdown. 54
2.3 Properties of the datasets. 57
2.4 Features of the global AIS dataset(provided by IHS). Due to this data being pre-fused by IHS Market, these are a combined set of variables that are normally across multiple AIS message types (see Table B.1. Items denoted with a dagger (†) text denotes the fields found in dynamic messages such as types 1, 2, and 3. Items marked with an asterisk (*) denote semi-static and static information predominantly found in type 5 messages. The items marked with a double dagger (‡) refer to the primary identifiers used by AIS (MMSI), and WRS (IMO). 66
2.5 Unique Identifier Analysis 66

3.1 Results of 20 scenarios each simulated over S runs of n time steps with 1 true vessel, v additional vessels, maximum of b additional objects, and p probability of measurements generated by the true vessel. The results SIAP completeness for all truths C_A, completeness for the true vessel C_T, GOSPA distance d. 89

4.1 A list of reflagging events presented to a human operator 107
4.2 An example set of stationary positions 110
4.3 An example set of stationary positions with details of their nearest ports. This can be visualised to show stopped location and nearest port as shown in Figure 4.5. 110

5.1 The number of detected outliers per ship type using the proposed approach with models trained using the AIS data from different ship types (rows) and tested on AIS data from ships of different ship types (columns). 117

5.2 Results of 10 scenarios each simulated over S runs with R regions, up to c change points, for n length time series. The r_s provides the mean and standard deviation of the Spearman's rank correlation coefficient for the posterior ranking described in Section 2.4 and r'_s denotes the mean and standard deviation of the Spearman's rank correlation coefficient for a random ordered ranking. 126

B.1 AIS Message Types 147
B.2 AIS Message Payload Variables 148
B.3 Navigation Status 152
B.4 Rate of Turn 152
B.5 Ship Types 153
B.5 Ship Types 155

C.1 Acronyms 156

Chapter 1

Introduction

Tracking, data fusion, machine learning and other forms of data science are equally applicable to many domains including the financial markets, logistics and security. It is through the application of such techniques that new algorithms are developed.

Maritime Situational Awareness (MSA) is the effective understanding of anything associated with the maritime domain that could impact the security, safety, economy, or environment [21]. In recent years, the requirement for MSA has increased due to the likes of terrorism [29], smuggling activities, piracy [56], protection of undersea cables [96], countering illegal fishing [62, 130] and illegal immigration [56].

The generation of an efficient maritime situational awareness picture of all maritime activity over an area of interest requires a mixture of surveillance systems, algorithms to support multi-source fusion and tools to support analysis [77]. A national maritime sovereignty picture includes all activity within the 200 nautical miles Exclusive Economic Zone (EEZ) (depicted in Figure 2.13).

Continuous tracking of all maritime activity by a single sensor is not sufficient and often not feasible. A single sensor cannot monitor everything that happens in the surveillance area. However there exists large data collection networks collecting and databasing information from many sensors. Therefore, to generate sufficient MSA the system needs to take advantage of the available data sources to construct a comprehensive maritime picture.

The monitoring of vast sea areas is a difficult and time consuming task for human operators trying to establish full Maritime Domain Awareness (MDA) [114]. This is due to the large amounts of heterogeneous data from multiple sources and the difficulties in detecting anomalous behaviour from normal maritime activities. The ability to automate

the detection of unusual activities would help decision makers efficiently monitor the ongoing activities in the surveillance area.

Anomaly detection is one of the enabling techniques for MDA. There are various anomaly detection techniques available that accomplish the maritime surveillance goals. Data-driven anomaly detection approaches find anomalous behaviour by constructing a model from normal data and calculating the deviation from that model. However, relying only on data-driven approaches for surveillance systems is not sufficient due to the lack of user involvement in the detection process [113]. More importantly, some of the suspicious behaviours of interest to human operators are activity specific and not directly observable, for example, a fishing vessel transiting the Atlantic is anomalous whereas a cargo vessel doing the same thing would not be anomalous. Therefore, it is impossible to find all types of anomalies by using data-driven approaches. Maritime domain experts have the required knowledge and experience for detecting maritime anomalies. Including the domain expert's knowledge about the suspicious activities of interest in the anomaly detection process that can result in an improved methodology.

The use of different data sources from open and closed data sources will influence the detection of activities of interest. These sources contain information about vessels, cargo, crew, etc. The closed data sources are only accessible to maritime authorities, such as the National Maritime Information Centre (NMIC) [36, 102, 133, 132], coastguard and law enforcement agencies. Some data are available online and freely (not necessarily free) accessible and reusable to the public [87]. There are also commercial suppliers [37]. Examples include; port authorities that publish their vessel traffic data and their facilities information online. In addition, there are many online communities such as blogs, forums and social media.

1.1 Thesis Overview

This thesis is focused on the application of data science and signal processing techniques to the maritime security environment.

The thesis explores different techniques to improve maritime situational awareness and describes data available to monitor global shipping, highlights several challenges faced when analysing such information, and documents the tools and techniques developed to collect, process and analyse this information to improve behavioural understanding and highlight events and behaviours.

Chapter 2 provides a review of the existing theory for probability density functions, Bayes theorem, state estimation and tracking. It also includes the theory of text analytics and change point detection. This chapter ends with a short description of the types of maritime data and sensors concluding with a description of the Automatic Identification System (AIS) and the datasets used in the remainder of the thesis.

Chapter 3 focuses on developing a method to handle the disambiguation of multiple vessels reporting on a single MMSI using a multiple target tracker. The research within this chapter utilises the Single Integrated Air Picture (SIAP) and the (Generalised) Optimal Sub-Pattern Assignment ((G)OSPA) metrics to assess the performance of the tracker using simulations of a set of known behaviours from the datasets.

Chapter 4 gives a review of analyses that focus on individual vessels.

Chapter 5 extends Chapter 4 by exploring the analyses that focus on aggregating the vessels to give a picture of the larger behaviours and areas of interest. This focusses on splitting the geospatial area into a set of regions. These regions are used to apply text analytics techniques (Latent Dirichlet Allocation (LDA) and Mixture of Unigrams (MoU)) and change point detection to the tracklets output from the tracker (Chapter 3) and the track joiner (Chapter 4).

The remainder of the thesis provides a summary and recommendations of areas that could be explored further.

1.2 Novel Contributions of this Thesis

- Using a multiple target tracker to disambiguate vessels sharing the same MMSI.
- Using track stitching to detect the probability of a vessel switching MMSIs.
- The use of a quad tree oriented abstraction of the geometry into regions for text analytic analysis.
- Change point detection on the region data to prioritise areas of high probability of a change occurring over multiple regions.

Subsequent to this work, these methods have been developed further and been demonstrated to Dstl and the NMIC as part of the Track Analytics project (see Appendix A), where this work focussed on the early algorithmic research task. The goal of Track Analytics was about how to convert the research into a product that can be demonstrated as part of

an integrated system downstream. The larger project team's work was about raising the Technology Readiness Levels (TRL) [45] of the algorithms and applying them to data in a particular software context, which is not described in this thesis.

1.3 Publications

The work described in Chapter 3, regarding the disambiguation of vessels sharing the same MMSI, was presented as a poster at the Defence and Security Doctorial Symposium 2018 [144]. The research in Sections 2.3.2.1 and 5.1.1.1, applying a MoU model to detect ship type from geospatial region abstracted tracks, was presented at the Sensor Data Fusion Symposium 2019 and published as [148].

Chapter 2

Background

This chapter provides an overview of the existing theories, topics and data covered and used in the rest of the thesis. It covers an overview of state estimation and tracking, track stitching, change point detection and text analytics techniques. The chapter also covers different maritime data types and specifically a detailed description of the Automatic Identification System (AIS) and the three AIS datasets used throughout the thesis.

2.1 State Estimation and Tracking

This section introduces the concepts of state estimation and tracking which form the main aspect of inferring the state of multiple objects over time. It begins with an overview of Bayes theorem, the Chapman-Kolmogorov equation, and the assumption of Gaussian noise.

2.1.1 Bayesian statistics

Given two events A and B, the conditional probability of A given that B is true is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

Bayes theorem [9] is used to update the probabilities after obtaining new data. $P(A|B)$ is defined as the posterior distribution, $P(A)$ represents the prior knowledge of A, $P(B|A)$ is the likelihood function which denotes the probability of B given that A is true and it quantifies to the extent that the data B supports the proposition proposed by the prior

knowledge A, and $P(B)$ is the marginal distribution which describes the set of all possible outcomes of the data.

The central limit theorem states that as the sample size, n of a distribution increases, the distribution of the sample mean approaches a normal distribution [115]. Therefore, noise associated with the stochastic processes can be represented by Gaussian noise.

2.1.2 State Estimation

The state of an object is the set of all possible variables describing the object. The state estimate is the subset of these variables that are being used in the estimation of the state over time. The state estimate is calculated from measurements of the true state of the object.

This section introduces the elements required to estimate the state over time.

State estimation uses a recursive process which takes the estimate, $p(\boldsymbol{x}_{t-1})$, from a previous time step, $t-1$, and predicts forward, $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, to the current time step, t. A measurement, $\boldsymbol{z}_t$, is supplied to the prediction to update the estimate, $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$. This follows a Markov process and is summarised by the following equations;

The Chapman-Kolmogorov equation [106] equates to the prediction step;

$$p\left(\boldsymbol{x}_t|\boldsymbol{z}_{1:t-1}\right) = \int_{-\infty}^{\infty} p\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right) p\left(\boldsymbol{x}_{t-1}|\boldsymbol{z}_{1:t-1}\right) d\boldsymbol{x}_{k-1} \tag{2.2}$$

and the Bayes' rule (adapting equation 2.1) equates to the update step;

$$p\left(\boldsymbol{x}_t|\boldsymbol{x}_{1:t}\right) = \frac{p\left(\boldsymbol{z}_t|\boldsymbol{x}_t\right) p\left(\boldsymbol{x}_k|\boldsymbol{z}_{1:t-1}\right)}{p\left(\boldsymbol{z}_t\right)} \tag{2.3}$$

where $p\left(\boldsymbol{x}_t|\boldsymbol{z}_{1:t-1}\right)$ is the probability of the current state given all measurements up to time step, $t-1$, and $p\left(\boldsymbol{x}_t|\boldsymbol{z}_{1:t}\right)$ is the probability of the current state given all measurements up to time step t.

2.1.2.1 Transition Equation

The state at the current time can be described as a process applied to the state at a previous time;

$$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{\omega}_t) \tag{2.4}$$

where $f_t(\cdot)$ is the transition function and $\boldsymbol{\omega}_t$ is the noise of the system. This noise takes into account that the motion model defined by the transition function might not be correct.

When the noise, $\boldsymbol{\omega}_t$ is additive, equation 2.4 can be simplified to

$$\boldsymbol{x}_t = F_t(\boldsymbol{x}_{t-1}) + \boldsymbol{\omega}_t \tag{2.5}$$

2.1.2.2 Measurement Equation

The measurement can be defined in terms of the state as follows;

$$\boldsymbol{z}_t = h_t(\boldsymbol{x}_t, \boldsymbol{\nu}_t) \tag{2.6}$$

where $h_t(\cdot)$ denotes the measurement function that translates the dimensions of the state space into that of the measurement space, and $\boldsymbol{\nu}_t$ is the measurement error.

Similarly, to equation 2.5, when the measurement noise is assumed to be additive, the measurement equation can be simplified to

$$\boldsymbol{z}_t = h_t(\boldsymbol{x}_t) + \boldsymbol{\nu}_t \tag{2.7}$$

where H_k is the measurement matrix. The measurement matrix selects the parts of the target state and converts it into the measurement space.

$$\boldsymbol{z}_t = \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1,t} \\ \dot{x}_{1,t} \\ x_{2,t} \\ \dot{x}_{2,t} \end{bmatrix} + \begin{bmatrix} \nu_{1,t} \\ \nu_{2,t} \end{bmatrix} \tag{2.8}$$

which produces

$$z_{1,t} = x_{1,t} + \nu_{1,t} \tag{2.9}$$

$$z_{2,t} = x_{2,t} + \nu_{2,t} \tag{2.10}$$

2.1.2.3 The Kalman Filter

The Kalman Filter (KF) is a recursive filter the produces an optimal minimum mean-squared error (MMSE) estimate of the current state of a system when the state follows a linear dynamic model with Gaussian noise [70]. The Kalman filter [70] is a popular method for state estimation in various domains such as agriculture [1], medicine [39], navigation [80] and finance [2, 76, 109, 141]. This section gives an overview of the Kalman filter algorithm (additional introductions to Kalman filters can be found in [5, 93, 124]).

A discrete time linear system can be described in the following state space form

$$\boldsymbol{X}_{t+1|t} = F\boldsymbol{X}_t + Q \tag{2.11}$$

where $\boldsymbol{X}_t$ is the state vector describing the state at the time step t, F is the transition model that defines the known motion of the linear system that is being observing. Q is the process noise. The measurement model can be described by the following equation;

$$\boldsymbol{Z}_{t+1} = H\boldsymbol{X}_t + R \tag{2.12}$$

where $\boldsymbol{Z}_{t+1}$ is the sensor measurement, H is the measurement matrix that converts the state space into measurement space. R is the measurement noise.

The Kalman filter is initialised with an initial state estimate $\hat{\boldsymbol{X}}_0$ and an initial covariance estimate S_0 for the time step $t = 0$.

Thus, the state estimate can be calculated at the time step $k+1$, $\hat{\boldsymbol{X}}_{t+1|t}$, based on the estimated state at time t, $\hat{\boldsymbol{X}}_t$, using the equation;

$$\hat{\boldsymbol{X}}_{t+1|t} = F\hat{\boldsymbol{X}}_t + Q \tag{2.13}$$

Similarly, the state error covariance at t is S_t and the update equation is

$$S_{t+1|t} = FS_tF^T + Q \tag{2.14}$$

where $S_{t+1|t}$ is the predicted state error covariance and Q is the covariance of the process noise. The Kalman gain, K is defined as

$$K = S_{t+1|t}H^T \left(\left|HS_{t+1|t}H^T\right| + R\right)^{-1} \tag{2.15}$$

where R is the covariance of the measurement noise. Once the Kalman gain is calculated, the update the state estimate, $\hat{\boldsymbol{X}}_{t+1}$, and the state error covariance, S_{t+1}.

$$\hat{\boldsymbol{X}}_{t+1|t+1} = \hat{\boldsymbol{X}}_{t+1|t} + K\varepsilon \tag{2.16}$$

where ε is the innovation defined as

$$\varepsilon = \boldsymbol{Z}_{t+1} - \left[H\hat{\boldsymbol{X}}_{t+1|t}\right] \tag{2.17}$$

and

$$S_{t+1|t+1} = [I - KH]\, S_{t+1|t} \tag{2.18}$$

.

2.1.2.4 Motion Models

The Taylor series expansion (see appendix E.1) is used to define the linear approximation of the transition relation of the state,

$$\begin{aligned} \boldsymbol{x}_{k+\tau} &= \boldsymbol{x}_{k+\tau} + \dot{\boldsymbol{x}}_{k+\tau}\tau + \ddot{\boldsymbol{x}}_{k+\tau}\frac{\tau^2}{2!} + \dots && (2.19) \\ \dot{\boldsymbol{x}}_{k+\tau} &= \dot{\boldsymbol{x}}_{k+\tau} + \ddot{\boldsymbol{x}}_{k+\tau}\tau + \dots && (2.20) \\ \ddot{\boldsymbol{x}}_{k+\tau} &= \ddot{\boldsymbol{x}}_{k+\tau} + \dots && (2.21) \end{aligned}$$

The approximation (neglecting higher order terms of the Taylor expansion) in matrix notation is

$$\begin{bmatrix} \boldsymbol{x}_{k+\tau} \\ \dot{\boldsymbol{x}}_{k+\tau} \\ \ddot{\boldsymbol{x}}_{k+\tau} \end{bmatrix} = \begin{bmatrix} 1 & \tau & \frac{\tau}{2!} \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \dot{\boldsymbol{x}}_k \\ \ddot{\boldsymbol{x}}_k \end{bmatrix} \quad (2.22)$$

The transition model is an example of a first order Markov model. The first order Markov model assumes that the state at time t can be derived from only the previous target state at time $t-1$. No further history of the state is required.

We can use the Taylor series expansion (see appendix E.1) to define the transition relation of the state,

$$\begin{aligned} \boldsymbol{x}_{k+\tau} &= \boldsymbol{x}_{t+\tau} + \dot{\boldsymbol{x}}_{t+\tau}\tau + \ddot{\boldsymbol{x}}_{t+\tau}\frac{\tau^2}{2!} + \dots && (2.23) \\ \dot{\boldsymbol{x}}_{t+\tau} &= \dot{\boldsymbol{x}}_{t+\tau} + \ddot{\boldsymbol{x}}_{t+\tau}\tau + \dots && (2.24) \\ \ddot{\boldsymbol{x}}_{t+\tau} &= \ddot{\boldsymbol{x}}_{t+\tau} + \dots \end{aligned}$$

We can then rewrite this as an approximation (neglecting higher order terms of the Taylor expansion) in matrix notation

$$\begin{bmatrix} \boldsymbol{x}_{k+\tau} \\ \dot{\boldsymbol{x}}_{k+\tau} \\ \ddot{\boldsymbol{x}}_{k+\tau} \end{bmatrix} = \begin{bmatrix} 1 & \tau & \frac{\tau}{2!} \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \dot{\boldsymbol{x}}_k \\ \ddot{\boldsymbol{x}}_k \end{bmatrix} \quad (2.25)$$

The transition function takes the form of a matrix in equation 2.11. The matrix form, in this particular case (), there are three degrees of motion; position, velocity, and acceleration.

The transition equation can be extended using the Taylor series expansion to higher order derivatives of position, for example, jerk.

$$\begin{bmatrix} 1 & \Delta_t & \frac{\Delta_t}{2!} \\ 0 & 1 & \Delta_t \\ 0 & 0 & 1 \end{bmatrix} \tag{2.26}$$

Equation 2.26 is used to derive the characteristics of motion. Firstly, if a vessel is stationary, the velocity component $\Delta_t = 0$. Thus, the matrix can be simplified to the constant position motion model (also known as a random walk or Brownian motion).

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.27}$$

The constant velocity model can be calculated in the same way. This time keeping the Δ_t.

$$\begin{bmatrix} 1 & \Delta_t \\ 0 & 1 \end{bmatrix} \tag{2.28}$$

The issue which can be encountered is that over time the covariance of a motion model can grow significantly (and in the case of vessels, the uncertainty can soon grow to the size of an ocean). One method to overcome this continual increase is to throttle the growth. This can be accomplished using a mean reverting process. A mean reverting process is defined as a process that its growth scaling decays the further from the mean it gets. An example of a mean reverting process is the Ornstein-Uhlenbeck Process [131].

$$F_t = \begin{bmatrix} 1 & \frac{1}{K}\left(1 - \mathrm{e}^{-K\Delta_t}\right) \\ 0 & \mathrm{e}^{-K\Delta_t} \end{bmatrix} \tag{2.29}$$

$$Q_t = q \cdot \begin{bmatrix} \frac{\Delta_t - \frac{2}{K}\left(1-\mathrm{e}^{-K\Delta_t}\right) + \frac{1}{2K}\left(1-\mathrm{e}^{-K\Delta_t}\right)}{K^2} & \frac{\frac{1}{K}\left(1-\mathrm{e}^{-K\Delta_t}\right) - \frac{1}{2K}\left(1-\mathrm{e}^{-2K\Delta_t}\right)}{K} \\ \frac{\frac{1}{K}\left(1-\mathrm{e}^{-K\Delta_t}\right) - \frac{1}{2K}\left(1-\mathrm{e}^{-2K\Delta_t}\right)}{K} & \frac{1-\mathrm{e}^{-2K\Delta_t}}{2K} \end{bmatrix} \tag{2.30}$$

Here K is our scaling of the decay.

As previously described, the Kalman filter is optimal for estimating states of systems

with linear dynamics, linear measurements, and Gaussian noise. Many real-world problems do not fall into this requirement for a linear application and require solving nonlinear systems whether in the dynamics or the measurements. As such, advances such as the Extended Kalman filter (EKF) [78, 127] and Unscented Kalman filter (UKF) [66, 67, 68, 69] extend the Kalman filter to be able to be used for non-linear problems.

2.1.3 Tracking

Target tracking focusses on dynamic estimation, by describing the real world in terms that can be arranged into mathematical formulae. Observations are taken and compared to the assumed target behaviour. The key important thing is the ground truth is only accessed through the sensor and measurement model. The mathematical models are relied on to be sufficient and mathematically tractable.

The object's state can be estimated using measurements from a sensor. For a single object moving at a constant velocity, a sensor in the same space can measure the position of that object to some degree of accuracy.

Typically, a sensor will find multiple objects in a space (of which some could be of interest), each of these objects could be following different trajectories to our original object.

A sensor detects an object, how the sensor detects an object and what they detect can be different for different types of sensors. For example, there are multiple possible properties a sensor can measure, such as position, velocity, size, etc. As well as the detections themselves the sensor will be able to detect at varying degrees of accuracy in, for example, time, and environment.

The compounding of the type of measurement and the detection methodology can result in a sensor that will detect clutter (false positives) as well as the measurements of the object.

The previous section outlined the process of tracking a single target using state estimation. This section will introduce the concept of track management and data association related to managing the tracking of multiple targets from a single sensor.

Single target tracking is concerned with whether a measurement is in line with the predicted state or not (an outlier). The assumption single target tracking makes is that a measurement was either generated by the vessel or it was not generated by the vessel, described as clutter. As such this produced a binary option for our Kalman filter to processes. The framework required to expand the problem to assume there is more than

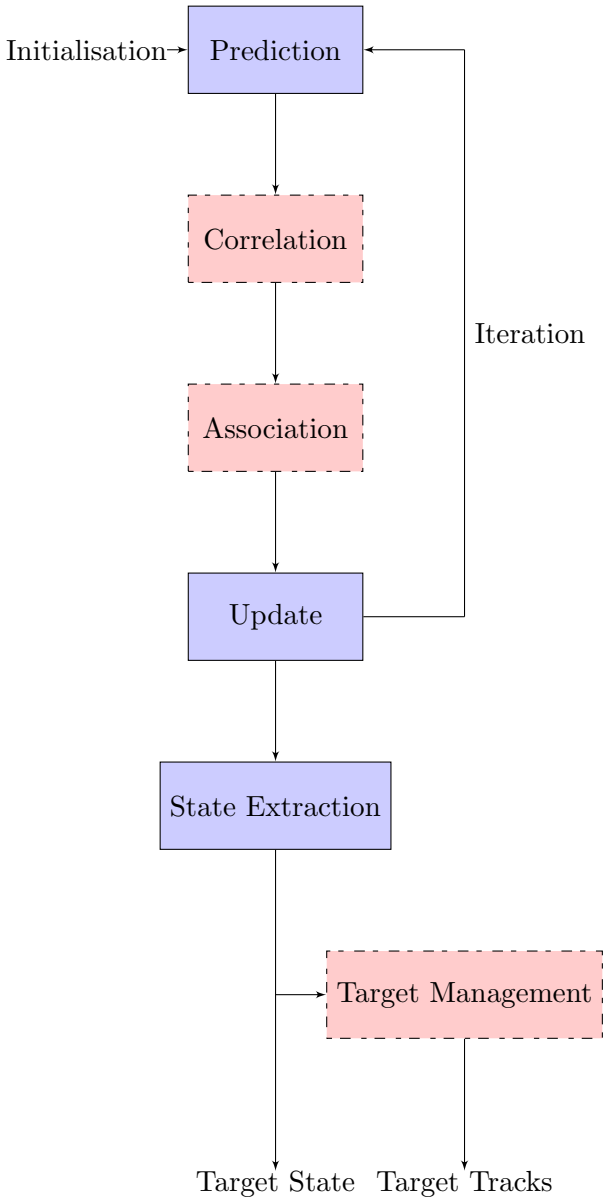


Figure 2.1: The Multi-Target Tracking Pipeline. The blue boxes represent the state estimation components discussed in Section 2.1.2 and the red boxes represent the track management and data association components associated with multiple target tracking discussed in Section 2.1.3.

one vessel is now introduced (see Figure 2.1).

Track Management

The method of deleting a given track is dependent on if a track has not had an associated observation for a certain length of time at which point it can be removed from the set of active tracks.

There are several methods of an existence function that can delete active tracks. Firstly, by setting a hard threshold of t time steps before an active track is removed. This function is only suitable when considering low numbers of active tracks in the tracker but becomes limiting as soon as the number of active tracks (potential vessel tracks) increases (e.g., > 1) as the number of tracks, ℓ, increases, the threshold per track (of t time steps) becomes $\frac{t}{\ell}$. An alternative version of the t threshold is to use a function that considers how many tracks exist within the tracker. The function is $t\ell$ where ℓ is the number of active tracks in the tracker.

These methods are suitable for datasets that have strict time steps or snapshot data like the global dataset but they are not suitable for data received at a native rate which do not arrive at uniform time intervals, such as the North Atlantic and Merseyside datasets where the time steps get converted to time durations.

The multiple target tracker is used to split the observations originating from a single MMSI number into corresponding tracks for each vessel using that MMSI number.

The construction of the multiple target tracker begins with each new observation being compared to the set of existing tracks within the tracker. If the observation is an outlier for all tracks, a new track is initiated from the observation. If the observation is an inlier for one or more tracks, the closest track is chosen. The closest track is updated with the new measurement while the remaining tracks are propagated forward to the new time step.

A track which has propagated for a certain number of time steps without an observation, is removed. This means that the potential vessel tracks that have not been seen for a certain length of time are removed. Managing the deletion of active tracks within the tracker is important to keep the computational cost low and to not allow the error covariances to get so large that they cause problems for associating observations with the correct tracks.

In this problem space there are N possible vessels (defined by their broadcast identification) and as such each identification number needs a tracker. There are also n vessels (broadcasting on a given identification number) a measurement could have originated from; therefore, a system is needed to manage the assignment of measurements to each of the corresponding vessel trackers.

Model Type	Missing Detections	Clutter	Best filter
Linear	No missing detections	No clutter	Kalman Filter (with predict/update of state and covariance)
Linear	Missing detections	No clutter	Kalman Filter (with predict and update of only covariance (track crossing))
Linear	Missing detections	With clutter	Probabilistic Data Association
Non-linear			EKF/UKF/CKF/PF

Table 2.1: Summary of tracking filters and the cases where they are best applied [125]

The previous sections in this chapter have described the components needed to construct a multi-target tracker for use with the AIS data set.

This section goes through the assumptions regarding AIS data (from Chapter 1 and existing algorithms and provides examples where the method is successful and where at least one assumption fails. This leads to the final modified method used for the rest of this study.

Figures 3.11, 3.12 and 3.13 clearly demonstrate the requirement for a multi-target tracker, using a state estimation filter, in this case, a Kalman filter, to determine inliers and outliers is not sufficient.

It is clear that this only works with the assumption that there is only one vessel reporting on a MMSI number and its associated noise. When considering that there may be more than one vessel reporting on a single MMSI, then a state estimation filter that classifies the observations as “inlier” (the vessel) or “outlier” (the noise) is not sufficient. By applying a multiple target tracker using that manages multiple state estimation filters allows all potential vessels reporting on the same MMSI number to be tracked.

There is a wide plethora of tracking solutions that include the Kalman filter family (standard Kalman filter (KF) [70], extended Kalman filter (EKF) [78, 127] and unscented Kalman filter (UKF) [66, 67, 68, 69]), cases for multiple measurements, such as the Probabilistic Data Association Filter (PDAF) [7, 4] and the Joint Probabilistic Data Association Filter (JPDAF) [6, 4] for multiple objects, and using the alternative random set framework, the Probability Hypothesis Density Filter (PHDF) [84].

Table 2.1 summarises the common tracking algorithms, their strengths and where they are best applied.

As tracking vessels using AIS is a linear problem, with occasional missing reports and

no clutter the Kalman Filter was chosen for this study.

2.1.3.1 Tracking individual targets

This section describes the components of a Kalman filter and its applicability to tracking through AIS data.

The symbology used throughout this section is summarised here:

k: the current time step
$k-1$: the previous time step
Δ_k: the time between $k-1$ and k.
$\boldsymbol{x} \in \boldsymbol{X}_k$: the set of true target states at time k (position and velocity)
$\boldsymbol{z} \in \boldsymbol{Z}_k$: the set of sensor measurements collected at time k.
$\hat{\boldsymbol{x}} \in \hat{\boldsymbol{X}}_k$: the set of estimated target states, an output of the multiple target tracking algorithm, at time k.

2.1.3.2 Measurement Space

For a multi-dimensional measurement space, a single measurement can be represented by the vector $\boldsymbol{z}$. Examples of measurement sensors and their associated measurement vector are;

- 2-dimensional space position measurement sensor (only record a target's position in x and y)

$$\boldsymbol{z} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{2.31}$$

- Passive sonar (only detect the bearing from a platform heading);

$$z = \phi \tag{2.32}$$

- Active sonar (detect the range to target from the sensor and the bearing.)

$$\boldsymbol{z} = \begin{bmatrix} r_k \\ \phi_k \end{bmatrix} \tag{2.33}$$

- Active radar (detect the range to target from the sensor, the bearing, and the elevation from the horizontal)

$$\boldsymbol{z} = \begin{bmatrix} r_k \\ \phi_k \\ \theta_k \end{bmatrix} \tag{2.34}$$

The important thing to notice is that a measurement vector can contain variables with different measurement units.

2.1.3.3 Target State Space

The state space of the target is similar to that of the measurement space but is likely to be higher in dimensionality and in a different coordinate frame. For example, from the sonar and radar examples in the section above, the measurements are being recorded in the polar coordinate frame, while the state space could be in the Cartesian coordinate frame. There are several convenient Cartesian representations of the world, for example a tangential plane with origin at a particular point in the Earth's surface, local flat earth. Issues in using this include the further from the origin you travel the further along the tangential plane you travel thus further away from the true position following the curvature of the earth. Earth-Centred Earth Fixed provides a 3-dimensional description of the world (a world made entirely from Lego blocks). The complexity of using this model space is that any movement of an object has to be in three dimensions.

Each 2-dimensional representation of the world fails to capture the true nature of the world. WGS84 [121, 32] and UTM [135, 122] are useful.

Examples of the state vector are;

- A state vector with 2 position coordinates;

$$\boldsymbol{x} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{2.35}$$

- A state vector with 2 position and 2 velocity components;

$$\boldsymbol{x} = \begin{bmatrix} x_k \\ \dot{x}_k \\ y_k \\ \dot{y}_k \end{bmatrix} \tag{2.36}$$

- A state vector with position, velocity, and acceleration components;

$$\boldsymbol{x} = \begin{bmatrix} x_k \\ \dot{x}_k \\ \ddot{x}_k \\ y_k \\ \dot{y}_k \\ \ddot{y}_k \end{bmatrix} \tag{2.37}$$

2.1.3.4 Measurement Equation

Assuming there is some noise in our measurements and states, the target states are represented by Gaussian distributions such that a "noisy" measurement or state can be defined as

$$\mathcal{N}(\mu, \sigma^2) \tag{2.38}$$

Where μ is the mean of the Gaussian distribution and σ^2 is the variance of the Gaussian distribution.

The generalised measurement can be described in terms of the state and a noise term

$$\boldsymbol{z}_k = h_k(\boldsymbol{x}_k, \boldsymbol{\nu}_k) \tag{2.39}$$

where k denotes time of the measurement, h_k is the measurement function that translates the dimensions of the state space to that of the measurement space, and $\boldsymbol{\nu}_k$ is the measurement error. The measurement equation (2.39) allows different ways of combining the state and the noise, e.g., multiplicative noise. Most commonly in target tracking additive

noise is required and the measurement function is applied to solely the state. This allows the measurement equation to be linear

$$\boldsymbol{z}_k = H_k(\boldsymbol{x}_k) + \boldsymbol{\nu}_k \tag{2.40}$$

where H_k is the measurement matrix. The measurement matrix selects the parts of the target state and converts it into the measurement space.

$$\boldsymbol{z}_k = \begin{bmatrix} z_{1,k} \\ z_{2,k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1,k} \\ \dot{x}_{1,k} \\ x_{2,k} \\ \dot{x}_{2,k} \end{bmatrix} + \begin{bmatrix} \nu_{1,k} \\ \nu_{2,k} \end{bmatrix} \tag{2.41}$$

which produces

$$z_{1,k} = x_{1,k} + \nu_{1,k} \tag{2.42}$$

$$z_{2,k} = x_{2,k} + \nu_{2,k} \tag{2.43}$$

2.1.3.5 Measurement Noise

The measurement noise is modelled with a zero mean multivariate Gaussian distribution and can assume that the components of the distribution are independent.

2.2 Track Stitching

This section introduces the concept of track stitching. When the output of a multiple target tracker produces track breakages; this method can re-join those broken track segments.

Track stitching as a post processing task applied to the output of a tracker. The stitching of two tracklets is simply predicting the last state of a tracklet forward to the start of the next tracklet [136]. The task of data association arises when there are multiple tracklets that meet the criteria of joining multiple tracklets. The joining of tracks in this case form an assignment problem. There are solutions using graph theory applications [26], using

network flow [19, 64]. The aim of the assignment is to find a perfect matching of tracklet pairs with a minimum cost.

The results produced by the disambiguation process in Chapter 3 provide a selection of tracklets. There are many reasons why a ship track could be incomplete including:

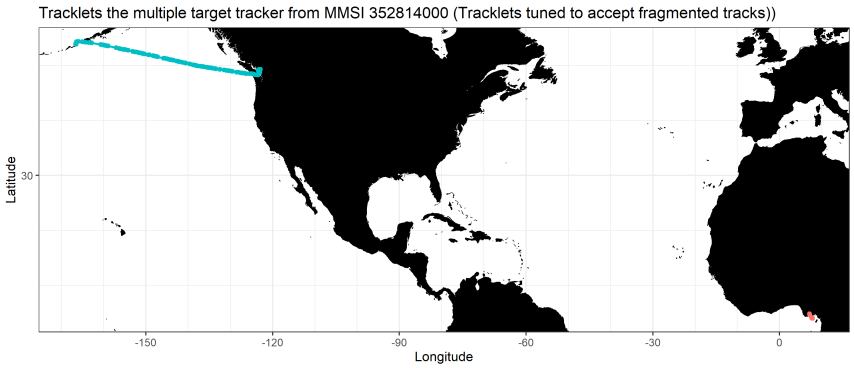
- A single vessel that had sailed in areas of poor AIS coverage
- Large update times for a single vessel being detected from space
- Two or more vessels broadcasting the same MMSI value and in close proximity
- Occasional erroneous position reports for a single vessel
- A single vessel changing its registration details (MMSI and call sign), for example with changing of ownership of re-registering to avoid a country's stricter maritime regulations. This is a special form or track stitching.

Vessel reflagging occurs infrequently so it is rare to find examples in smaller extracts of AIS. There are analogous problems and situations with different datasets (specifically not in the open literature) where the identity (ID) for a given object changes over time (with some regularity). The processing and analysis of AIS data therefore has a need to allow for reflagged vessels. More interestingly the domain of radio frequency detection (see Section 2.6.1.3) has a need to keep track of an object that will regularly (and frequently) change their identifier as the vessel swaps use of its 3GHz and 9GHz radars.

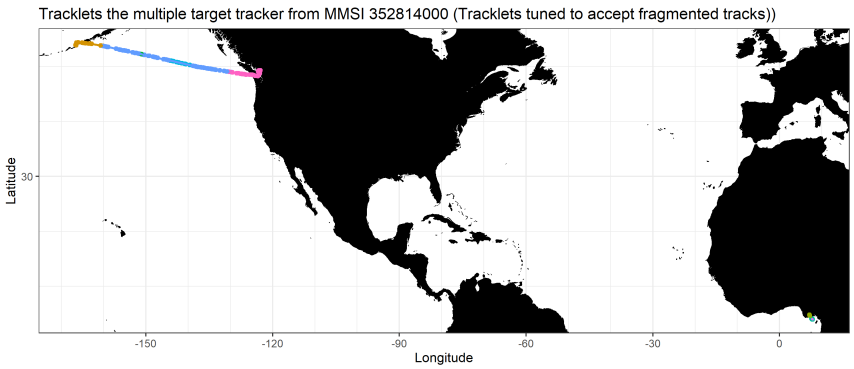
This section describes track stitching techniques and shows results from putting these in practice. Using the output of the multiple target tracker, described in Chapter 3 .

2.2.1 Forward Prediction

The output of the disambiguation shown in Chapter 3 is a set of tracklets which each are formed by likely paths vessels may have taken. The ability to estimate how many vessels are reporting on a given MMSI is required (it should always and only be 1, despite the numerous examples shown in Chapters 1 and 3). The bank of Kalman filters within the multiple target tracker have been tuned in such a way to allow voyages to essentially break the chain of a true trajectory for the duration of the dataset. Figure 2.2 shows the difference between the tuning parameters of the disambiguation such that a vessel is completely tracked (Figure 2.2a) versus a trajectory that is fragmented into daily segments (Figure 2.2b).

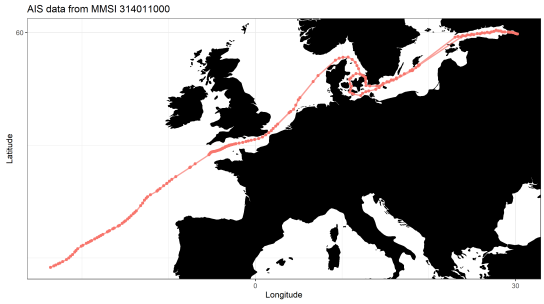


(a) The daily voyages of a vessel as a single tracklet.

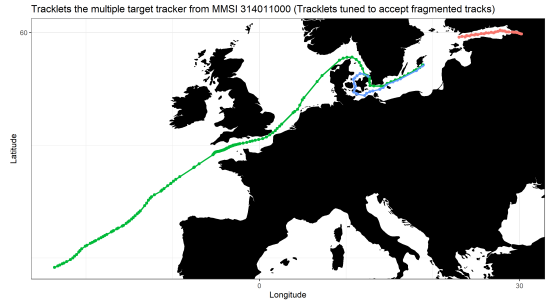


(b) The daily voyages of a vessel as individual tracklets.

Figure 2.2: An example where the entire vessel trajectory is a single tracklet compared to the vessel trajectory being split over multiple tracklets.



(a) The result of tuning the MTT such that a single tracklet for the MMSI in Figure 2.2a on a different MMSI.



(b) The result of other MMSIs when using the MTT is tuned such that each daily voyage is an individual tracklet.

Figure 2.3: The results of overfitting from Figure 2.2 applied to a different MMSI

So, if the multiple target tracker (MTT) had been tuned to accept this single example, it would be over fitting the data and as a result let two vessels sharing a MMSI be recorded as a single tracklet (the entire idea of producing tracklets was to stop this happening).

Now, assuming all MMSIs produced the output of Figure 2.2a, the number of tracklets per MMSI could then be counted to estimate the number of vessels using a given MMSI. Since the method results in overfitting (and producing the Figure 2.3a diagram for all vessels is labour intensive (tuning each MTT for each MMSI)), a method that can adapt the output of Figures 2.2b and 2.3b is needed to produce a join a string of tracklets as a single vessel and leave out the tracklets that are likely to have originated from a different vessel.

A method of joining follows that of network flow examples in computer vision. Here the combinatorics of the tracklet decision tree are produced of potential joins and use the state estimate and covariance to propagate a tracklet reduced to a start point and an end point forward and the gating (extended past the disambiguation limit of the deleter from the multiple target tracker) to see if there are viable tracklet pairs to be joined.

The tree or set of possible paths can be thought of from the initial time (the start of the dataset) to a termination time (the end of the dataset). If the end point of a tracklet overlaps the start point of another tracklet, there is a very low probability that these tracklets could actually be a continuation of a true track. If a tracklet begins a good long time after another tracklet ends, the probability of them joining would also be very low.

- All tracklets are assumed that they must start after the start of the dataset, such that

each tracklet can be joined with the start of the dataset. Similarly, the same applies to the end of the tracklets can be joined to the end of the dataset. Every complete track begins at the start of the dataset, proceeds through a subset of tracklets and concludes at the end of the dataset.

- Any start of a tracklet is assumed that it has to have come from an end of another tracklet and the end of that tracklet must then go on to the start of another tracklet.
- A tracklet cannot, mid-way through its lifespan, decide to be another tracklet.

For two tracklets to be joined, they must be temporally and geospatially nearby to each other (the end of first tracklet to the start of the second tracklet).

There are two use cases for this methodology, firstly as described above, when a tracklet from the given MMSI is likely to have been "accidentally" split (as Figures 2.2 and 2.3 demonstrate), and secondly, where a vessel has changed its MMSI (potentially by reflagging). If a vessel has changed its MMSI the two tracklets to join will be from different MMSIs and, the combined-tracklets (from first step, the auto join method) and the remaining (leftover) tracklets from each MMSI with tracklets from every other MMSI need to be compared.

Figure 2.4 provides an understanding of the process of disambiguation happening on a MMSI by MMSI scale. Each filter only deals with a single MMSI (each of which are the multiple target trackers from Chapter 3 to allow for instances where more than one vessel can be using the same MMSI). Each of these filters in the sensor bank then send their information to the centralised track stitching engine that calculates the probability of associating two tracklets and the probability that this is a reflagging event.

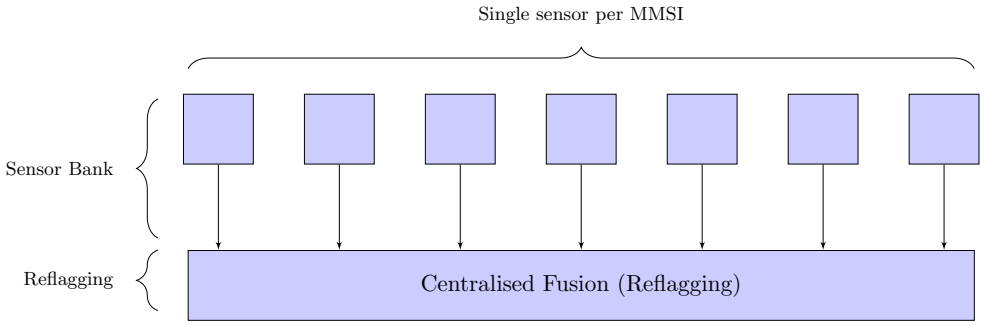
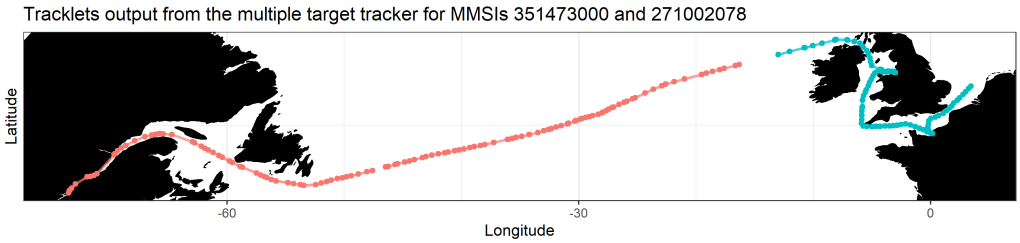


Figure 2.4: Fusion from sensor disambiguation to centralised reflagging.

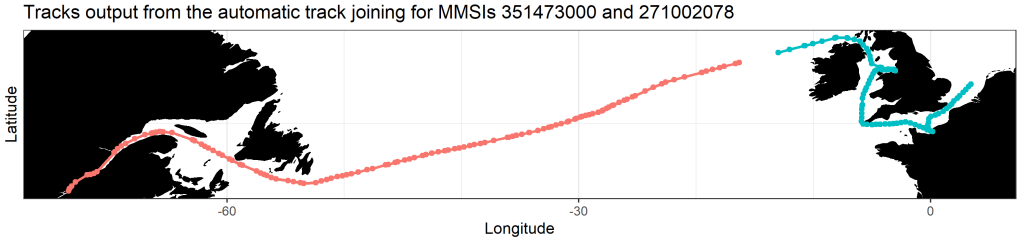
Since it is required to join only tracklets that are temporally close and spatially close,

this drastically reduces the computational cost.

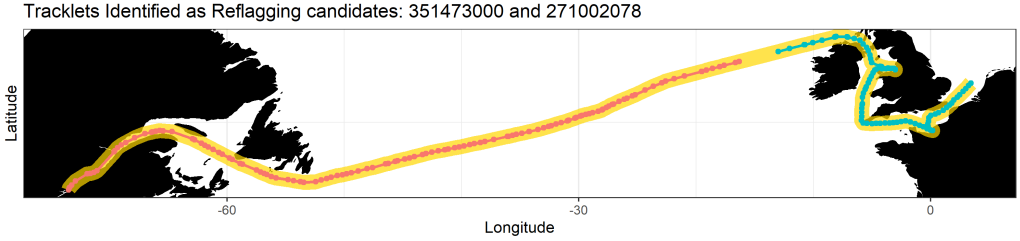
As there is a higher probability that two tracklets from the same MMSI are actually part of the same track rather than from another MMSI, the automatic tracklet join method is run first (comparing tracklets within the same MMSI) and then the reflagging method (comparing tracklets from different MMSIs). Figure 2.5 provides a full workflow of the automatic tracklet join and the reflagging method.



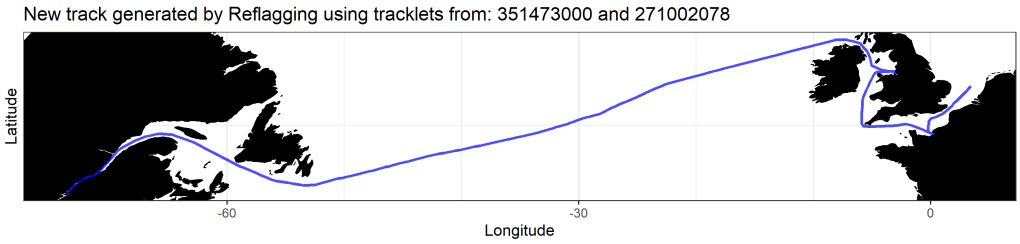
(a) Tracklets produced by the disambiguation.



(b) The automatic tracklet join combines tracklets from the same MMSI.



(c) The reflagging method provides a human operator a set of candidate reflagging events.



(d) The human operator has the ability to approve and assign a new custom identity to reflagging vessels.

Figure 2.5: The process of using the automatic tracklet join followed by the reflagging method on a set of disambiguated tracklets.

2.2.2 The Rauch-Tung-Striebel Smoother

The Kalman filter introduced in Section 2.1 provides a filter. The filter is a forward operator. All our tracklets are able to predict the next time step. To improve the cost associated with the joining of a tracklet pair correctly, the tracklets are additionally tracked backwards. To predict backwards, such that the end of a tracklet could predict forwards in time and the start of the next tracklet could predict backwards to verify the join.

The Rauch-Tung-Striebel smoother [110] is a backwards smoother. It firstly performs a forward pass, which is identical to that of a Kalman filter. The filtered state estimates ($\hat{\boldsymbol{x}}_{k|k-1}$ and $\hat{\boldsymbol{x}}_{k|k}$) and covariances ($\hat{\boldsymbol{P}}_{k|k-1}$ and $\hat{\boldsymbol{P}}_{k|k}$) from the Kalman filter are stored to be used in the backward pass, the smoother.

The smoother processes the state estimates and covariances starting from the last time step and working its way backwards to the first observation's state estimate and covariance.

The result of this smoothing process provides the possibility to predict the tracklet backwards in time. This means that Rauch-Tung-Striebel smoother can utilise the state and covariance over the prior Gaussian used initiating the track in the multiple target tracker) which provides a more meaningful prediction of the state at the previous time step.

The following information for each tracklet is obtained;

- Tracklet ID,
- MMSI,
- Start and end time,
- Start and end state and covariance estimates,

and the combinatoric decision tree of all possible tracklet combinations to predict forward from the end of a tracklet and predict backward from the start of a tracklet to calculate the probability that they could have come from the same trajectory.

The track stitching is implemented in Sections 3.2 and 4.1 where the performance is assessed and then applied to the joining of tracklets within a given MMSI, denoting track breakages generated by the multiple target tracker and the case where a vessel has deliberately changed its MMSI number resulting a tracklet ending in one MMSI and a tracklet starting in another MMSI.

2.3 Text Analytics

This section provides a description of the geography abstraction that generates a set of geospatial regions and an overview of the text analytics algorithms, Latent Dirichlet Allocation (LDA) and the Mixture-of-Unigram (MoU) models, that take a set of documents infer a set of data derived topics for each document.

Prior to doing any behaviour analysis, the geospatial area containing the position reports is separated into smaller sub-regions.

These regions are used to extract behavioural characteristics of vessels and of the regions themselves.

2.3.1 Adaptive Grid and Geospatial Clustering

An adaptive grid based on the quad-tree [116] is used to split a large region into a set of smaller grids based on their observation count such that for a selection of geospatial points, this information is represented as region specific "characteristics" and group regions by similarity.

The adaptive grid method developed turns the density of geospatial points into a set of regions varying in size defined by a maximum threshold of geospatial points a region can contain. Using a quadtree [116] structured adaptive grid, if a region contains a count higher than the threshold, the region is further split into four sub regions along the mid-point of that region. This divides the data such that within the busy areas, where ship behaviours are diverse, a more precise representation of positions is obtained. Elsewhere, where ship density is lower, a coarser representation is obtained. Figure 2.6 shows the output of the adaptive grid on a UK focused subset of the global dataset. The adaptive grid processed the 1.6 million observations into 5,000 regions based on a threshold of 1,000 observations per region.

Once the grids were generated each observation in the dataset was allocated its associated grid cell ID. Each track was then converted from geospatial way-points to the grid cell ID that contains that way-point. The resultant tracks were a list of symbols identifying the grid cells the track passes through. Figure 2.7 gives an example of a track converted into a sequence of symbols.

It is these chains of symbols that are used to perform text analytic probabilistic analysis.

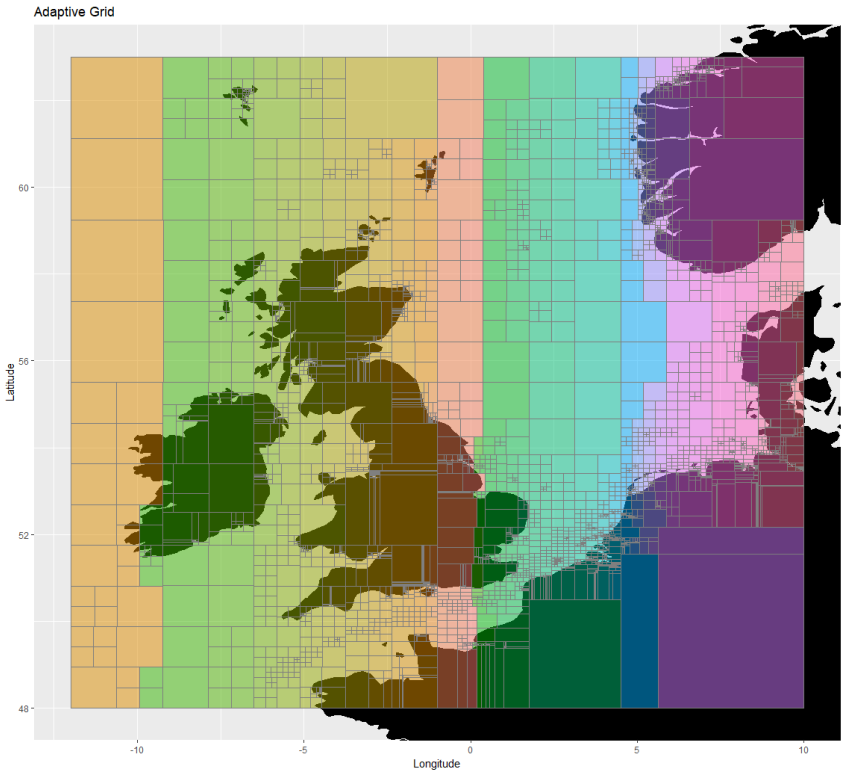


Figure 2.6: Adaptive grid applied to the UK subset of the global dataset.

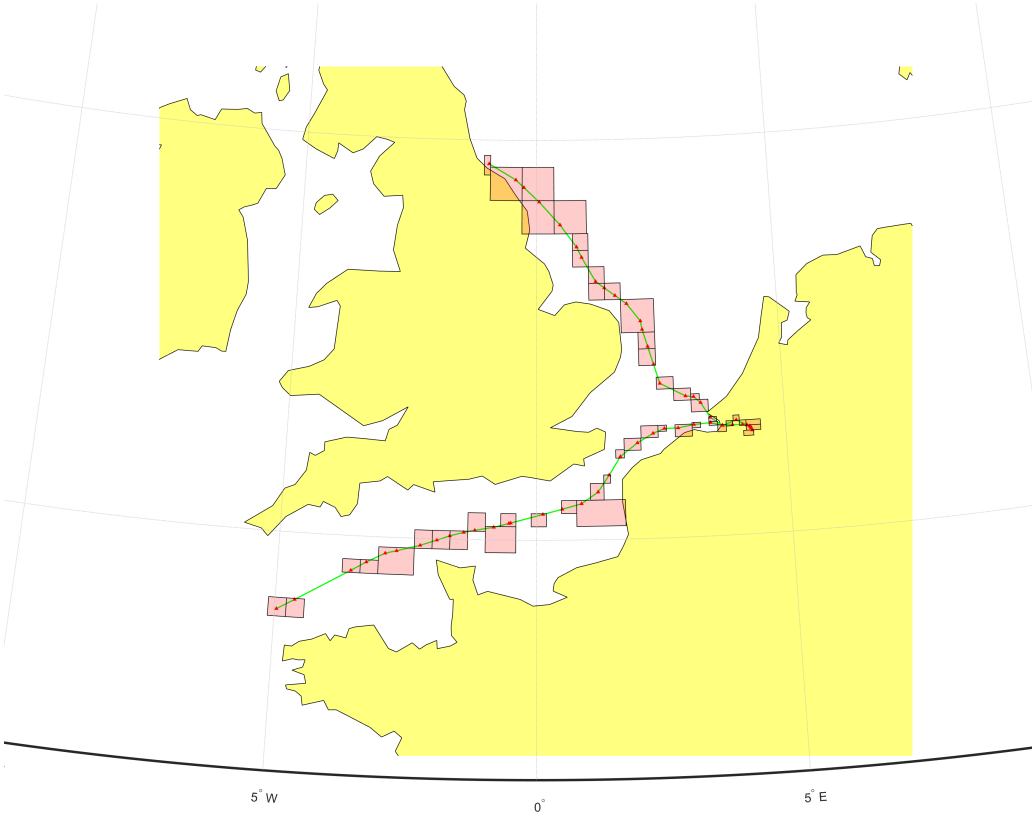


Figure 2.7: An example of how to convert a latitude-longitude track into a sequence of symbols. In this track, "$32 : 2, 50 : 10, \ldots$" is a part of the extracted data (i.e., document) and means that the ship stayed in region 32 for two AIS messages and stayed in region 50 for 10 AIS messages. Note that as the interval between two consecutive AIS messages is one hour, two AIS messages from the same region imply that the ship stayed in the region for about two hours.

2.3.2 Introduction to Probabilistic Topic Models

There are two widely developed algorithms: Mixture-of-Unigrams (MoU) [103] and Latent Dirichlet Allocation (LDA) [11], that are used extensively in text mining (and are applicable in any context involving symbolic features) where this approach is challenging to apply in practice. These algorithms assume that the documents' topics are sampled from a multinomial or Dirichlet distribution, and the words in documents are sampled from dictionaries described using multinomial distributions. For both models, the likelihood of a new data point can be measured by calculating the probability density function. Thresholding the likelihood for anomaly detection is feasible (e.g., [95, 140]). However, one of the advantages of both algorithms is that they consider an arbitrary length feature as input: the number of words in a document is not always fixed.

This section describes the MoU and LDA models to show how they work.

2.3.2.1 Mixture-of-Unigrams - *The Dirichlet Multinomial Mixture model*

The Dirichlet-Multinomial Mixture provides a model that can generate a document for a given topic and generate words from the topic. This is otherwise known as the Mixture-of-Unigrams model [148].

Assume that there are N training data, $X = x_{1:N}$, and an explicit prior[1], which can be thought of in terms of a pseudo-prior, i.e., as N_0 data, $X_0 = x^0_{1:N_0}$. The machine learning algorithm can be thought of estimating the value of a parameter, π. While it is often not a feature of the algorithm that is used, after training, the generative model can then be evaluated (i.e., the posterior predictive distribution) at a test point, x, as $p(x|\pi, X, X^0)$.

An alternative scenario is now considered such that there are $N_{0'}$ data in the pseudo-prior but the statistical moments of the two pseudo-priors ($X^0 = x^0_{1:N_0}$ and $X^{0'} = x^{0'}_{1:N^{0'}}$) are the same. For a fixed value of π, $p(x|\pi, X, X^0)$ can be calculated, along with $p(x|\pi, X, X^{0'})$.

An outlier is considered as a datum that is less similar to the training data than is expected. That datum must be (comparatively) more similar to the prior. Assuming $N_{0'} > N_0$, the problem of testing if x is an outlier can be posed as model selection between the two models that the statement $p(X^{0'}|\pi, X, x) + p(X^0|\pi, X, x) = 1$ exists. By using the extended form of Bayes theorem (considering two competing hypotheses and the law of

[1]The approach demands that the prior is explicit, but doesn't demand that it is informative

total probability), the probability of being an outlier can be derived as follows.

$$\begin{aligned} p(X^{0'}|\pi, X, x) &= \frac{p(x|\pi, X, X^{0'})p(X^{0'})}{p(x|\pi, X, X^0)p(X^0) + p(x|\pi, X, X^{0'})p(X^{0'})} \\ &= \frac{1}{1 + \frac{p(x|\pi,X,X^0)p(X^0)}{p(x|\pi,X,X^{0'})p(X^{0'})}} \end{aligned} \tag{2.44}$$

where $\frac{p(X^0)}{p(X^{0'})}$ is the prior odds-ratio for the two models.

$$p(\boldsymbol{w}|\pi, X, X^0) = \sum_z \int_{\beta_1} \dots \int_{\beta_T} p(z|\pi, X, X^0) \left[\prod_{t=1}^{T} p(\beta_t|\pi, X, X^0)\right] \left[\prod_{j=1}^{N} p(w_j|z, \beta_z)\right] d\beta_1 \dots \beta_T \tag{2.45}$$

$$\begin{aligned} &= \sum_z p(z|\pi, X, X^0) \left[\prod_{t=1}^{z-1} \underbrace{\int_{\beta_z} p(\beta_t|\pi, X, X^0) d\beta_t}_{=1}\right] \\ &\times \int_{\beta_z} p(\beta_z|\pi, X, X^0) \left[\prod_{j=1}^{N} p(w_j|z, \beta_z)\right] d\beta_z \left[\prod_{t=z+1}^{T} \underbrace{\int_{\beta_t} p(\beta_t|\pi, X, X^0) d\beta_t}_{=1}\right] \end{aligned} \tag{2.46}$$

$$= \sum_z \underbrace{p(z|\pi, X, X^0)}_{\text{Sample a topic}} \underbrace{\int_{\beta_z} \overbrace{p(\beta_z|\pi, X, X^0)}^{=Dir(\beta_z;\eta)} \left[\prod_{j=1}^{N} \overbrace{p(w_j|z, \beta_z)}^{=Mn(w_j;\beta_z)}\right] d\beta_z}_{\text{Analytic integral, procedure of sampling words: see (2.49)}} \tag{2.47}$$

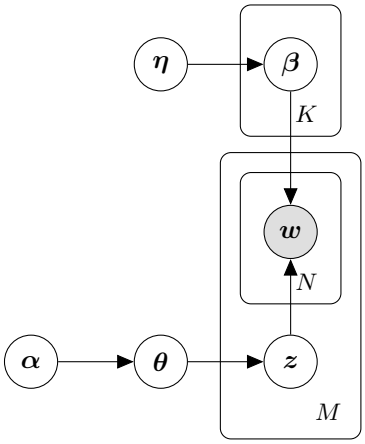


Figure 2.8: The graphical representation of the Mixture-of-Unigrams model. The circles mean samples drawn from a distribution. The boxes mean a number of replications. α and η are the parameters for Dirichlet Distributions from which the multinomial distributions are drawn. θ and β define multinomial distributions that represent the topic and word distributions, respectively. z and w are samples that are drawn from the corresponding multinomial distributions.

Specific instances of the approach for Mixture-of-unigrams

Although the MoU model [103] is not only applicable to text analysis, in this section, the following notations and terminologies is still used to aid intuitive understanding.

- **Word** (w) is the basic symbolic representation unit in the data. There are assumed to be V words in the vocabulary.
- **Document** ($\boldsymbol{w}$) includes a sequence of N words which is denoted by $\{w_1, \ldots, w_N\}$.
- **Corpus** (**D**) includes all the documents in the training set. The size of the corpus is denoted by M, such that $D = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$.
- **Topic** (z) is an attribute of document and each document has only one topic[2].
- **Topic Distribution** (θ) is a parameter of the entire corpus and follows a multinomial distribution. It is represented by a $K \times 1$ vector where K is the number of topics in the corpus and describes how often each topic occurs.

[2]The LDA differs from the MoU by considering each document to have a mixture of topics

- **Word Distribution per Topic** (β_z) indicates the parameter of a multinomial distribution of words assigned to each topic, z. It is a $K \times V$ matrix.

The MoU model (shown in Figure 2.8) generates a corpus using the following steps: For the corpus, sample the distribution of topics, θ, from the Dirichlet distribution, $Dir(\alpha)$; Sample the distribution of words per topic, β_z, from $Dir(\eta)$; For each document, sample a topic, z, based on θ and the number of words, N, using a Poisson distribution; For each word in the document, sample from the word probability, β_z, based on the sampled topic, z. The probability of a generated document is:

$$p(\boldsymbol{w}|\theta, \beta) = \sum_z p(z|\theta) \prod_{n=1}^{N} p(w_n|z, \beta) \tag{2.48}$$

Accordingly, the posterior predictive of an input (new) document can be calculated, (2.45)-(2.47), given the learnt model, π, the pseudo-prior, X^0, and training data, X.

The analytic integral in (2.47) is derived from the procedure of sampling word distributions per topic from a Dirichlet prior (the first term) and sampling words in the document given the topic assignments (the second term), such that it can be calculated using the conjugate relationship stated in (2.49)-(2.51).

$$\int_{\beta_z} Dir(\beta_z; \eta) \left[\prod_{j=1}^{N} Mn(w_j; \beta_z) \right] d\beta_z = p(w_{1:N}|\eta) \tag{2.49}$$

$$= p(w_1|\eta) \prod_{j=2}^{N} p(w_j|w_{1:(j-1)}, \eta) \tag{2.50}$$

$$= p_{D_i}(w_1|\pi, X, X^0) \prod_{j=2}^{N} p_{D_i}\left(w_j|\pi, X \bigcup w_{1:(j-1)}, X^0\right) \tag{2.51}$$

where $Dir(\beta_z; \eta)$ and $Mn(w; \beta_z)$ respectively denote the Dirichlet and Multinomial distributions. This is such that (2.51) is a function of the posterior predictive distribution, which for a Dirichlet-Multinomial model is:

$$p_{D_i}(x = k|\pi, X, X^0) = \frac{\alpha_k + n_k}{\sum_{j=1}^{N}(\alpha_j + n_j)} \tag{2.52}$$

where n_k is the count of the (training) data in the kth class; α_k is the corresponding (pseudo-)count from the pseudo-data comprising the pseudo-prior.

In summary, (2.47) can be calculated as a weighted sum of analytic integrals, one for each potential topic that the input document could be a member of. n_k is estimated during the training procedures of the MoU, and α_k can be manipulated, which is equivalent to X^0 in Section 2.3.2.1, to calculate the probability of generating this document using the models that are more similar to the training data or the prior.

A Mixture-of-Unigrams generates all the words in a given document from exactly one topic, z. This differs from the LDA model (below) where a single document can express multiple topics.

2.3.2.2 The Latent Dirichlet Allocation Model

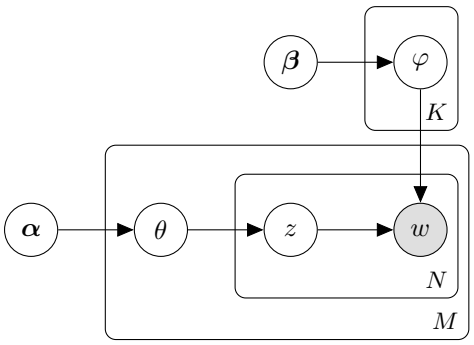


Figure 2.9: Graphical model representation of the Latent Dirichlet Allocation model

The corpus, depicted in Figure 2.9, contains M documents and each is a sequence of N words. Open circles are parameters (α, β, θ, φ) or latent variables (θ, z). The shaded circle is the observed word variable (w) and boxes (plates) represent replicates. The Dirichlet parameter, α, and topic-word matrix, β, are corpus-level parameters sampled once in the process of generating a corpus. The topic proportions, θ, is a document-level variable sampled from α once per document. The topic, z, is a word-level variable sampled from θ once for each word in a document. Formally, a K-topic LDA specifies a

two-level probabilistic process that generates a document as follows, (i) a K-dimensional vector, θ, is chosen from the distribution $p(\theta|\alpha)$, and (ii) words are sampled repeatedly from the document-specific mixture distribution, $p(w|\theta)$. Exact inference and parameter estimation involve calculating the posterior distribution on a document $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$. This is intractable because the latent variables are coupled via the edge between θ and z. The posterior can be approximated by computing the variational Dirichlet parameter θ and the variational multinomial parameter φ for each word in the document. The subscripts m, n, and k on a parameter (β, θ, φ) or variable (θ, z, w) donate the m-th document, n-th word and k-th topic respectively.

Note that the Dirichlet variable α is a distinct component of the probability model and not merely an expression of uncertainty about a parameter.

2.3.2.3 Specific instances of the approach for Latent Dirichlet Allocation

Although the LDA model [103] is typically applied to text corpora, this technique can be generalised to solve problems in other fields such as:

- Bioinformatics [147]
- Computer Vision [17]
- Social Network Analysis [146]

In this section the LDA is applied to behaviour analysis. The same notations and terminologies are used to aid intuitive understanding.

- **Word** (w) is the basic symbolic representation unit in the data. Assume that there are V words in the vocabulary.
- **Document** ($\boldsymbol{w}$) includes a sequence of N words which is denoted by $\{w_1, \ldots, w_N\}$.
- **Corpus** (**D**) includes all the documents in the training set. The size of the corpus is denoted by M, such that $D = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$.
- **Topic** (z) is an attribute of behaviour.
- **Topic Distribution** (θ) is a parameter of the entire corpus and follows a multinomial distribution. It is represented by a $K \times 1$ vector where K is the number of behaviours in the corpus and describes how often each behaviour occurs.

- **Word Distribution per Topic** (β_z) indicates the parameter of a multinomial distribution of behaviours assigned to each region, z. It is a $K \times V$ matrix.

These methods are applied to disambiguated tracks where the positional geography has been abstracted to geospatial regions in Section 5.1. The LDA model is applied to the abstract tracks to infer journey behaviours (Section 5.1.1) and the MoU model is allied to infer the vessel type (Section 5.1.1.1).

2.4 Change Point Detection

This section introduces the techniques for detection change points in a time series and defines a score based on the likelihood of a change occurring in the time series that allows multiple series being ranked by the probability of largest change.

This section proposes data derived intelligence such that the information gleamed from the process can improve operator workload. The method proposed prioritises operator effort in detecting geographical regions of interest by detecting changes in activity. Existing methods for an operator to search an area for a temporal based anomaly is by brute force. By using the count of vessels in a given region over time we can model the behaviour of vessels being in that geospatial region. Some regions will remain fairly constant, but some regions' vessel counts can change. It is possible to predict these changes and the severity of the change. This area of mathematics has been studied extensively in statistics and other domains of application.

Traditionally, multi-source fusion uses intelligence from other sources which can then be applied to the data. We define context-aware data fusion as intelligence derived from data. What we do is aggregate up the low level AIS data to generate hypotheses of context about geospatial regions. Change point detection is an example of this. Getting the aggregated count of ships in a region suddenly makes you think "Our intelligence source, consisting of AIS at scale, tells us there is a down surge of activity near Somalia today". This is an intelligence output but is derived from individual ships reporting their position and not being (or being) in a certain place.

Figure 2.10 displays data generated from a Poisson distribution with changing mean λ. This sample dataset is a typical example of change point data used throughout this section.

Change point detection is the calculation of certain points within a time series that denote a break in the previous defining parameters of the time series with a new set of

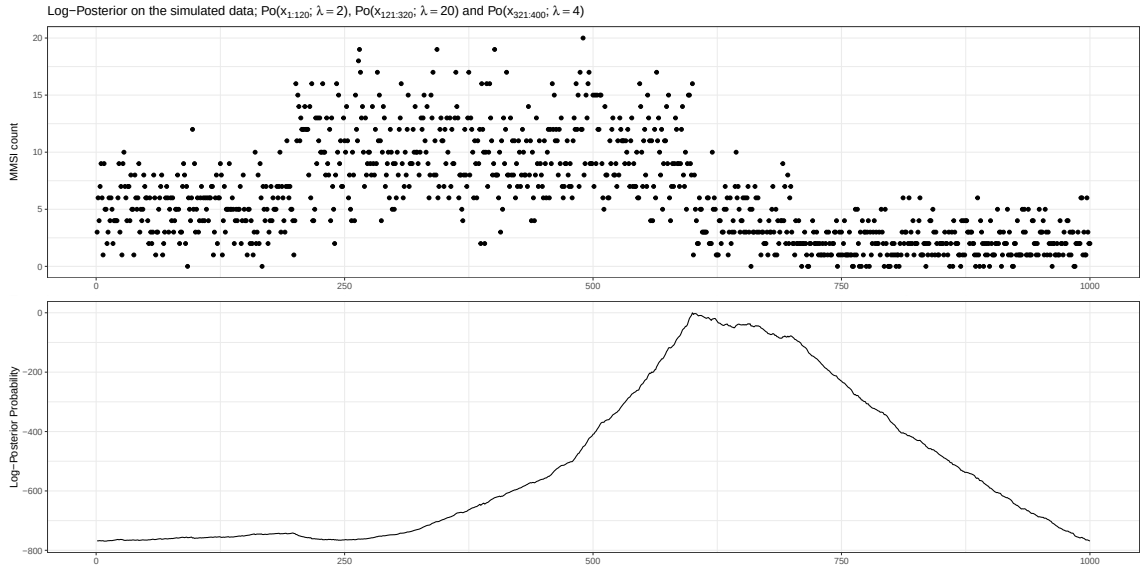


Figure 2.10: 400 observations of Poisson change point data x with abruptly changing mean. The points represent observations drawn from the underlying generating mean, λ with values 2, 20 and 4

parameters defining the series after the selected point. Change points are abrupt changes in the parameters of a sequence of observations. There are many change point detection processing tools that are either off-line or online. There are significant areas of interest in online frequentist methods; [105], [119], and Bayesian approaches are off-line (with some exceptions).

Change point detection focusses on the detection of a change point occurring in a time series of data whether it be a single change (At Most One Change, AMOC) [22], or change in the mean and(/or) variance [46, 71, 119], or a change in the regression model[22]. The literature focusses on extracting enough information out of a single time series to find the point at which a change occurred, the parameters of the series before the change point and the new parameters of the series after the change point. The focus of the following described method, aims to detect a series from a set of time series that contains a significant change in behaviour to warrant further inspection.

The aim of all these methods is to offer a detailed analysis of a single data sequence/time series with one or more change points. What we propose is the use of a negative binomial distribution method to analyse thousands of datasets (in this case, each dataset is the MMSI

count per hour for each region defined by the gridded geospatial regions (see Section 2.3.1). We assume there are two hypotheses on the data. H_0 such that all the samples are generated from a single Poisson distribution with a single parameter which is unknown and H_1 that all the data after a point are from a Poisson distribution with a different parameter. The result is the log probabilities of all points being from a different distribution, i.e., if we can collate all values H_j, we would end up with a complete log probability for each point being a change point. This analyses all datasets for a change to have occurred in that region and return a probability of a change within the dataset. The processed regions are then ranked by the likelihood of a change point.

2.4.1 Change Point Detection from Simulated Count Data

100 Monte Carlo simulations were performed to detect if a region contains any change points. For each test, a set of parameters were generated for the generation of the data: the number of distributions between change points, the parameter *lambda* for each distribution and the length of each distribution subset. We pass the simulated test to the change point algorithm which provides the log posterior for a change being detected.

We assume that the number of distributions, $D \sim \text{Po}(6)$, the value of λ for each distribution is sampled from the set $\{1, 2, \ldots, 20\}$, and the number of points per distribution are generated from $n_i \sim 1000\frac{n'}{\sum n'}$ where $n' \sim U(D; 0, 1)$.

Figure 2.11 presents the number of change points for each test against the detected number of change points from the detection algorithm. We can see that there is a positive correlation between the true quantity and the detected quantity. It can also be seen that the majority of tests detected fewer change points than the true number of change points. This can be described by the distance between consecutive mean counts. If, for example, the mean of two consecutive distributions are similar, then the probability of a change point being detected is lower than if there was a larger difference in the consecutive means, then the probability of a change point being detected is higher. This is illustrated in Figure 2.12. Figure 2.12 shows for similar consecutive means, the change point detected is lower than the detection probability for divergent consecutive means.

We can consider a maritime example which can describe these scenarios. If we have an area of the ocean in which fishing takes place and we have a geospatial region which may or may not be in the fishing area. If there are 24 vessels in the fishing area, then we can detect approximately 24 vessels in the geospatial region every hour. If another vessel entered the

fishing area, we would then be detecting approximately 25 vessels in the region every hour. This slight change in the count is noticeable but not a large change in behaviour. If all 25 vessels were to leave the fishing area, we would be detecting very few if any vessels in the region. This dramatic change in activity is important to detect.

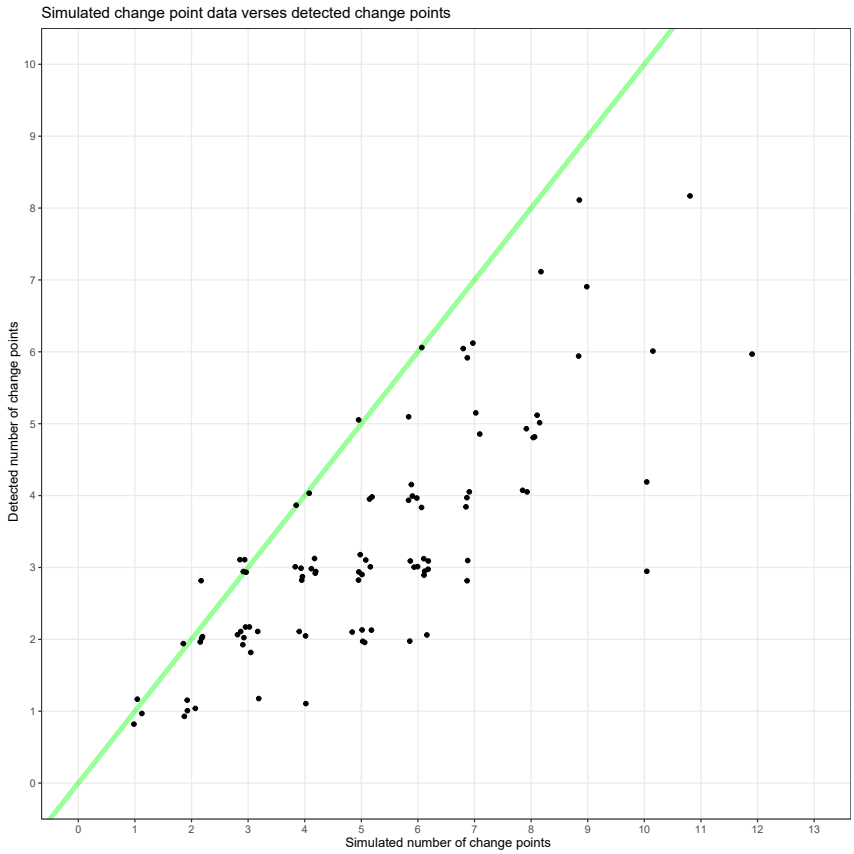


Figure 2.11: Simulation of change point data and the total number of changes detected by the algorithm for 100 tests. (Jitter applied to the integer data.)

The methods described here are utilised in Section 5.2 where the techniques are applied to the geospatial regions to detect changes over the set of regions.

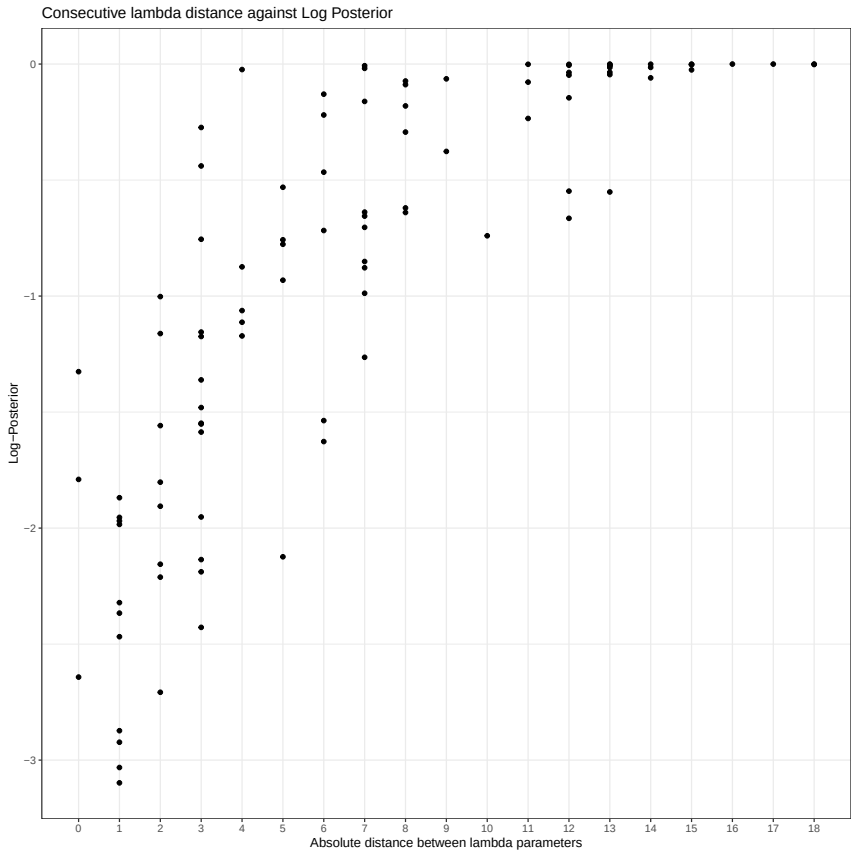


Figure 2.12: Simulation of change point data with a single change point and two differing distributions and the log posterior of the data produced by the algorithm for 100 tests. (Jitter applied to the integer data.)

2.5 The Maritime Challenge

The movement of goods by sea is a largely unregulated activity. The high seas are classed as international waters and as such no single government or jurisdiction has authority. Figure 2.13 shows the different areas; territorial waters, the exclusive economic zone and international waters. Figures 2.14 and 2.15 show the number of vessels reporting from that latitude and longitude against the percentage of ocean for each given latitude and longitude. There is an explicit freedom of movement.

There is some regulation, jointly developed, accepted, and implemented by member states of the International Maritime Organisation (IMO) [57], that cover aspects such

as Safety of Life at Sea (SOLAS) [51], curtail pollution and regulate the movement of hazardous materials. Consequently, much effort is needed to monitor and police these waters [41]. Governments and industry both have a requirement to monitor the oceans. The UK Strategy for Maritime Security "outlines the UK's approach to delivering maritime security at home and internationally" [100]. One of the strategies recommendations was the forming of the National Maritime Information Centre (NMIC) to improve cross-government collaboration and information sharing. Within the NMIC, information from many sources is collected and analysed to support government countering maritime threats.

Maritime threats include:

- Terrorism
- Human trafficking
- Smuggling [104]
  - Drugs
  - Tobacco products
  - Arms and ammunition
- Piracy
- Money laundering
- Breaches of UN sanctions
- Illegal fishing
- Damage to the environment
- Denial of freedom of navigation

In 2016 there were around 58,000 vessels in the world trading fleet, of size > 100 gross tonnes [34]. These perform many tasks including moving cargo, oil, and installing infrastructure. If you add the global fishing fleet, private vessels, and government owned vessels then this number is well in excess of 200,000 vessels. Detecting the threat vessel or group of vessels from such a large number, in such a loosely regulated arena, is clearly going to be challenging.

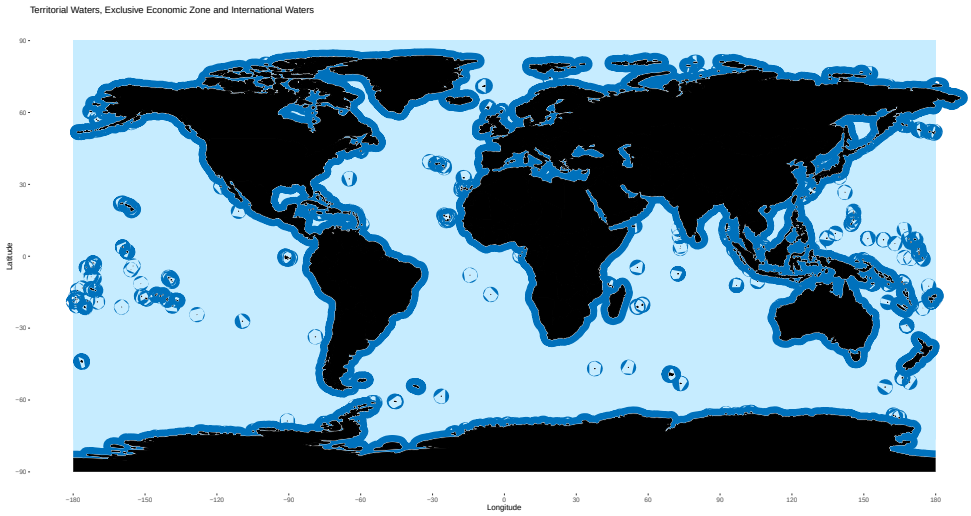


Figure 2.13: Visualisation of land and inland water (black), territorial waters [12nm] (white), exclusive economic zone [200nm] (dark blue), and international waters (light blue).

In 2000 the IMO introduced Regulation 19 of SOLAS Chapter 5 [53] requiring the installation of shipborne Automatic Identification System (AIS) to be fitted on most commercial ships 300 gross tonnage and above and passenger vessels carrying 12 passengers or more regardless of size [56]. This requirement was later amended and became a requirement for all commercial shipping over 300 gross tonnes by the end of 2004. Ships fitted with AIS broadcast their name, position, course and speed to other vessels and ground stations within their vicinity over a VHF broadcast. On board ship these messages are used to augment the radar display to improve awareness of other vessels intention and ultimately reduce the number of collisions at sea. On land these broadcasts are collected and aggregated to provide a situational awareness picture.

There are many challenges with processing global maritime information, especially as surveillance from space borne sensors improve.

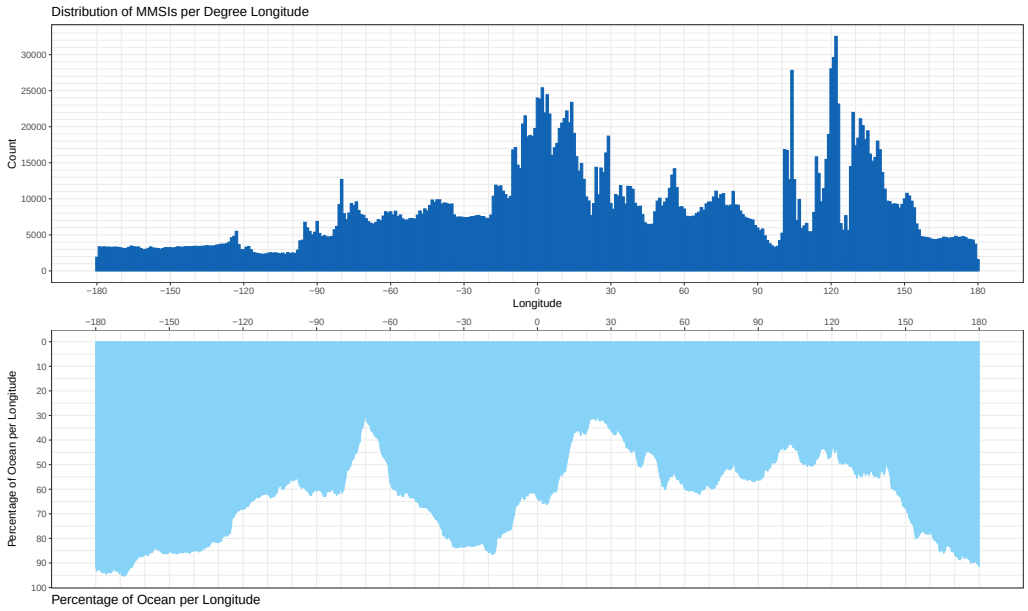


Figure 2.14: Distribution of vessels (unique MMSI numbers) per degree longitude (dark blue) against the percentage of ocean per longitude (light blue).

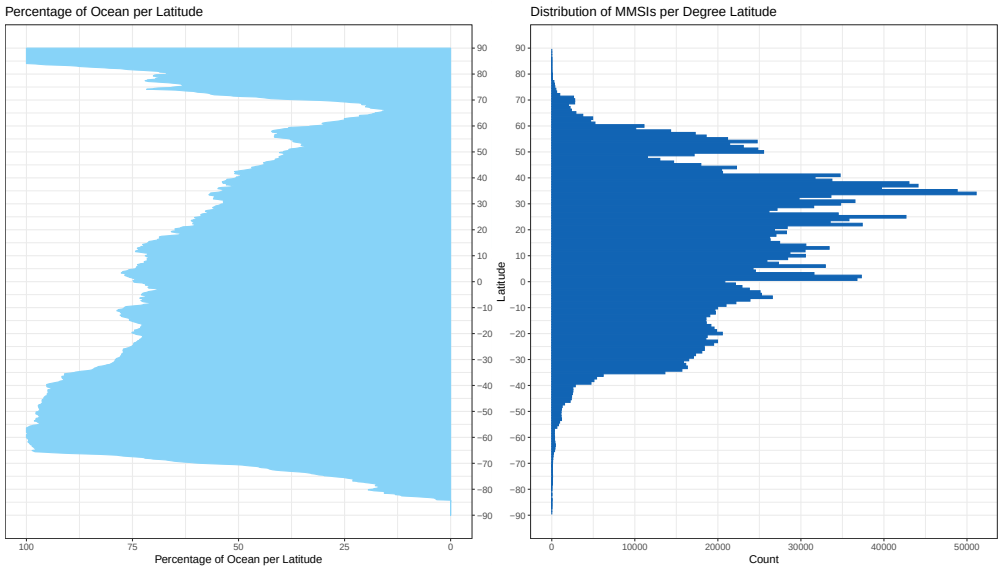


Figure 2.15: Distribution of vessels (unique MMSI numbers) per degree latitude (dark blue) against the percentage of ocean per latitude (light blue).

2.6 Maritime Data and Sensor Types

This chapter further introduces the maritime environment, lists a selection of the requirements and describes some of the information types available and the peculiarities and challenges in processing such data.

Due to its availability much of the research has used AIS information [61]. One of the challenges in using AIS is that a vessel's name and position are broadcast in different messages using a key field, the Mobile Maritime Service Identity (MMSI), to associate the reports. Although vessels should have a unique MMSI, there are occasions where multiple vessels are using the same value leading to ambiguities.

There are many types and sources of information that can be exploited to create a global maritime picture. These include open sharing of information, purchasing information from commercial suppliers and extracting information from websites. In general, these data types are divided into two classes: cooperative and non-cooperative.

Typically, surveillance problems are approached from fusing multiple sources of data with varying degrees of trustworthiness and accuracy.

This section provides a brief description of these information types to help the reader understand not just how data can be recorded but how it is derived, the scope of the information available and how it can be justified as trustworthy.

2.6.1 Non-Cooperative Data

Non-cooperative sensors are sensors that provide information independently to the vessel. The collection does not rely on the vessel in providing any information, i.e. the vessel is not required to cooperate and provide data for these sensors to detect them.

2.6.1.1 Radar

Radar is a method of radio detection which has the ability to detect the direction by azimuth and elevation, and range to a target.

2.6.1.2 Space-based Radar

Space-based radars have large/regional areas of coverage and are starting to be used more to support maritime surveillance. By using a technique called Synthetic Aperture Radar (SAR) it is possible to construct a two-dimensional image of a target region. 3-dimensional

reconstructions of landscapes can also be produced [72]. The SAR can provide a finer resolution than beam-scanning radars by using the motion of the antenna over the region of interest. SAR can be mounted on moving platforms such as aircraft and satellites. The SAR matches the distance it travels over a region to the time taken for the pulses to return to the antenna to create a synthetic aperture of the antenna. As a result, the larger the aperture (synthetic or otherwise), the higher the resolution of the image. This means SAR can create large resolution images with a relatively small physical antenna.

The main challenge with space-based radars is the re-visit time. For a single satellite it is often only possible to take one radar picture of an area near to the equator every day however as more space-based radars are coming on-line, the update rates are improving.

2.6.1.3 Radio Direction Finding

By using radio direction finding, it is possible to determine the bearing of a vessel from a headland. In the UK, the Coastguard has a number of radio direction finding antennas tuned to the VHF marine band and use these to determine the location of a vessel in distress at sea.

It is also possible to determine the location of a vessel by analysing the Radio Frequency (RF) signature of the vessel's radar. Using direction finding and time difference of arrival techniques, it is possible to measure and triangulate the direction from which an RF signal was transmitted and hence determine the location of the vessel. The commercial RF detection from space has been implemented by Hawkeye360 [8] and RF emitters being geolocated by a swarm of Uninhabited Aerial Vehicles (UAVs) [117].

2.6.2 Legislation

There are a number of maritime legislation and 'rules of the sea' [89]. Rules are defined as the systematic process of running a vessel at sea. Such rules define the use of shipping lanes in confined areas, environmental constraints such as the use of low sulphur oil in Northern Europe and the need to report the carrying of any hazardous materials. Port authorities also have local rules determining the maximum speed and the use of a pilot or tugs to assist with mooring.

This section describes some of the legislation applicable to the tracking of vessels at sea.

2.6.2.1 Safety of life at Sea (SOLAS)

These are a set of books for mariners that report restrictions, maybe due to cable laying, outline shipping lanes and other rules of the sea.

2.6.2.2 International Maritime Organisation

The International Maritime Organisation (IMO) is the United Nations specialised agency with responsibility for the safety of shipping and the prevention of marine and atmospheric pollution by ships [57].

The IMO issue each vessel with a unique IMO number that stays with the vessel for life. The use of the Automatic Identification System is also mandated under the IMO SOLAS legislation [55].

2.6.3 Databases

There are a number of commercial databases containing up-to-date information about vessels and maritime activities. Most of the information stored in these databases are static information about the vessel, for the vessel specifications/parts etc.

There are a number of suppliers of maritime information including:

- Marine Traffic . This database is accessible through the internet and access to most fields is free [79].
- IHS Markit [48].
- Lloyd's List Intelligence [83].
- Vessel Finder [137].

2.6.3.1 IHS Markit World Registry of Shipping

IHS Markit provide a web portal, named Sea Web, providing vessel, vessel ownership and other information [47]. It is regularly updated and provides the de facto ground truth for vessel specifications.

For off-line use the Royal Navy purchase an extract from this database named the World Register of Shipping (WRS). A copy of this database was kindly supplied to the University of Liverpool to support the Track Analytics project.

This is a great database of all ships on the ocean held by the IMO. The Registry of Ships has a whole host of information about the provenance of a ship, including ownership, insurance and types of cargo. This database was used to assist with the validation of ships in this thesis.

2.6.3.2 List of Ports/UN Locodes

The United Nations Code for Trade and Transport Locations, UN/LOCODE database is a huge database that contains uniquely identifiable information on all sorts of travel contexts and not just maritime. The database provides a set of ports, airports, rail networks, post offices, administrative districts, etc. and managed and maintained by the United Nations Economic Commission for Europe (UNECE) secretariat [134]. The aim of the UN/LOCODE database is to correctly identify ambiguous port names by creating a code based on the ISO 3166-1 country code [59] and three letters marking the port name.

Figure 2.17 shows that the use of UN/LOCODES can identify ports with ambiguous names. Despite this, the list of ports is not complete. Not all ports are included. The process for a port to be included in the database is to apply to the UNECE. That being said, there are issues with the integrity of port data in the database. There are examples where locations in latitude and longitude stored as decimal degrees have been entered as degrees, minutes, and seconds and vice versa, resulting in many ports in the database that are not on the coast. These are visible in Figures 2.16 and 2.17.

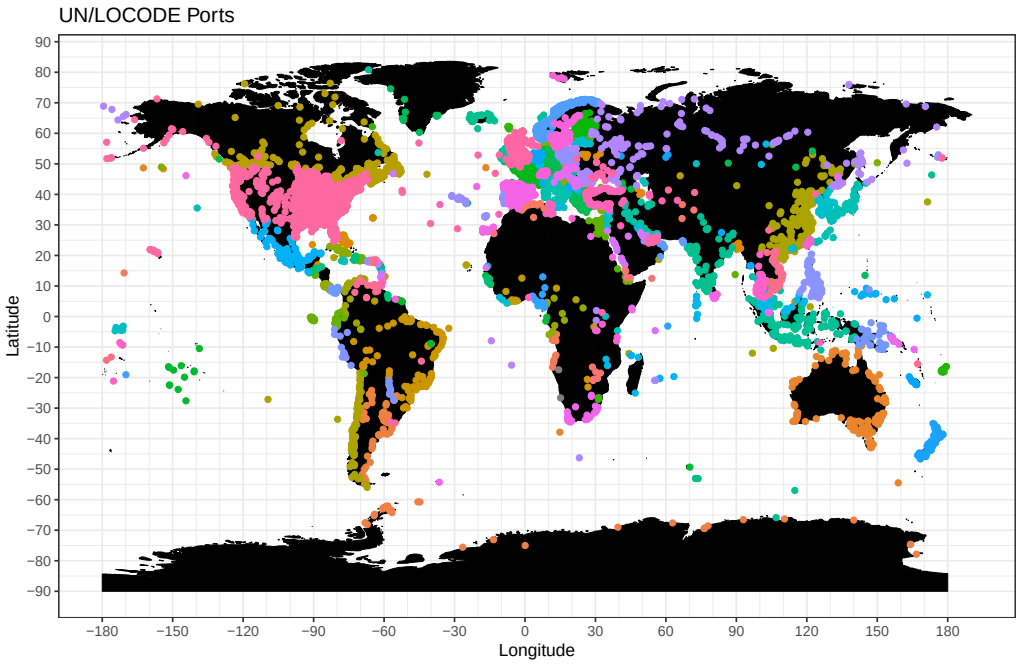


Figure 2.16: The ports defined in the United Nations Code for Trade and Transport Locations grouped by country.

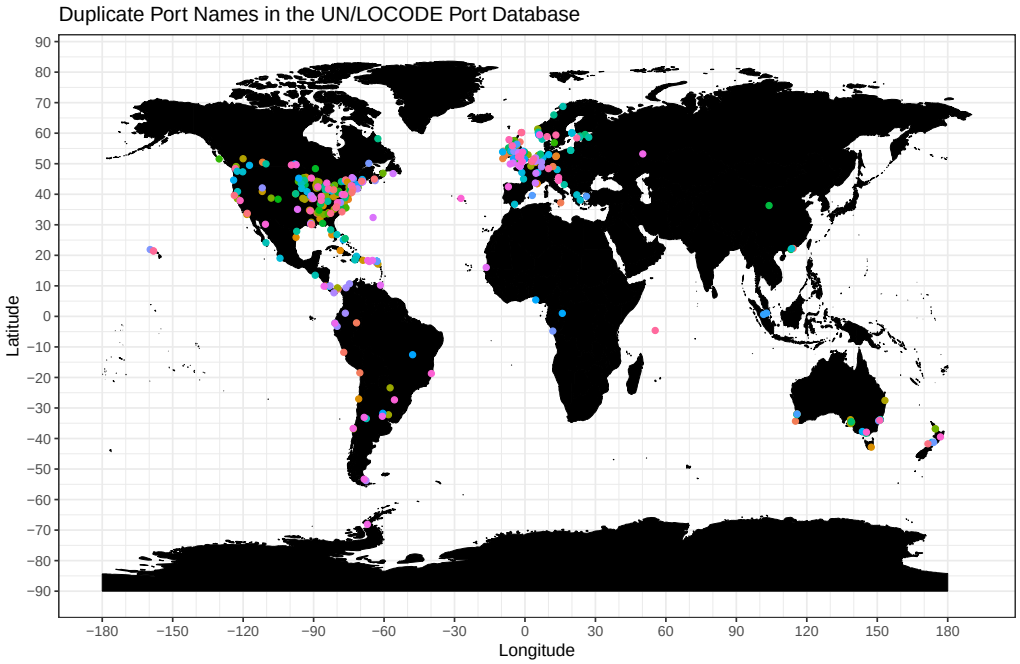


Figure 2.17: Duplicate port names in the United Nations Code for Trade and Transport Locations Database. The depicted 371 ambiguous ports outline the discrepancy of using a text box to name a location over using a system such as the UN/LOCODE as a destination for the Destination field used in the static AIS messages.

2.6.4 Cooperative Data

This section outlines the data that are provided by the vessel. The majority of information issued by a vessel is correct however there is the potential for a vessel to issue false information.

2.6.4.1 Automatic Identification System

The Automatic Identification System (AIS) was mandated by the IMO to limit the potential for collisions at sea. Messages are broadcast over the VHF radio band to adjacent ships containing a vessels name, contact details and destination. These messages are used to annotate the radar plot on a vessel.

The IMO legislation requires all vessels over 300 gross tonnes broadcast AIS messages. In addition, pleasure craft can also participate on a voluntary basis.

Many identities use AIS to develop a picture of maritime traffic. These pictures can be around a small area, such as a port, or by sharing information it is possible to build a comprehensive large area picture. Some companies have installed global networks of AIS receivers and sell this data to appropriate users.

More recently some commercial data providers have launched AIS receivers on satellites to observe shipping in mid ocean [13, 24].

Due to its availability AIS information was used to support the algorithm development in this thesis.

For a more detailed overview and description of the AIS data structures see Appendix B.

2.6.4.2 Long Range Tracking and Identification

The use of Long Range Tracking and Identification (LRIT) is mandated by the IMO under the SOLAS ch. V convention [53]. It became mandatory for all vessels over 500 Gross Tonnes in January 2009.

LRIT was designed to aid the tracking of vessels where monitoring using AIS is difficult. Examples include open ocean [20] and hostile territories such as the Gulf of Aden [58].

Ships using LRIT broadcast their location to a data centre owned or designated by the country of registration. These data centres then sell the information on to other requesting, and entitled, countries.

Entitled countries include:

- Vessels within 200NM of their territory (with some exceptions)
- Vessels declaring they are heading to their country
- Countries involved in the rescue of vessels in distress

Access to LRIT data is limited to governments and their agents and hence LRIT is not used in support of this thesis.

2.6.4.3 Vessel Monitoring System

The Vessel Monitoring System (VMS) is mandated by the European Union for the monitoring of fishing activities and management of fish stocks. Fishing vessels over 12 metres in length are mandated to report their location to their national fishing authority every 6 hours.

As the location of fishing grounds is often commercially sensitive, in many cases passed down from father to son, the VMS system uses private satellite-based communications to report their location.

Access to VMS data is limited to government only and hence VMS has not been used to support the work in this thesis.

2.6.4.4 Notice of Port Arrival

Under European Law vessels are required to give 24 hours notice of arrival into a European port.

2.7 Automatic Identification System

The Automatic Information System (AIS) is a communication system that broadcasts messages on the maritime Very High Frequency (VHF) band. AIS contains a communication protocol that determines the information to be transmitted and the equipment that utilises the protocol to send and receive messages.

AIS vessel tracking has been a significant development in navigation safety. Land-based AIS receivers used by port and safety authorities use AIS data to manage water transportation and reduce navigation accidents. Vessels travelling close to coastlines (40 nautical miles (on a good day this can be in excess of 150NM)) can be received by land-based AIS base stations. Lately, satellites have been launched to receive AIS data from all over the

world in order to extend the coverage area beyond what was possible with the exclusively land based stations [15]. This is a new development, which provides new opportunities for monitoring and analysing global ship traffic.

AIS enables an automatic exchange of information from the vessel to another source. The data that can be transferred includes static data, such as navigational and ship details, dynamic data, such as speed, and semi-static or voyage related information, such as draught, destination and Estimated Time of Arrival (ETA). The typical use of the system is to exchange information with near-by ships to avoid high risk situations such as collisions at sea. It is also used in traffic management between stations on shore and vessels at sea, specifically, for port management.

AIS messages contain information such as vessel locations, vessel names, Search and Rescue helicopters, virtual buoys and more. These messages are used to provide an augmentation of the radar-based collision avoidance system on board a vessel. As a result, AIS enabled vessels are provided with a clearer picture of their location and the position of nearby transmitting vessels. It should be noted that AIS is only mandated on large commercial vessels (SOLAS Class A vessels). The standard provides for smaller vessels and pleasure craft (SOLAS Class B vessels) however this is voluntary. Consequently, an AIS picture does not include all sea going vessels.

AIS is gathered by dedicated VHF receivers, these can be found onboard vessels, on buoys, on land and satellites [50]. The implementation and development of AIS was an international project with the IMO and the International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA) to name a few! The development was initiated in 1994 and the regulations about the use of AIS were amended in the International Convention for the Safety of Life at Sea (SOLAS) [15, 51, 142].

The guidelines were formalised with AIS requirements in SOLAS ch. V 2002 such that all ships of 300 gross tonnage and higher engaged in international voyages, cargo ships of 500 gross tonnage and higher not engaged on international voyages, as well as all passenger ships built after 2002, or operated after 2008, must have an AIS system installed [54].

2.7.1 Message Types

The International Telecommunications Union (ITU) have defined 27 different message types that can be broadcast over AIS [61] a description of these message types is given in annex B.

The AIS messages support both SOLAS Class A and Class B vessel types. Table 2.2

outlines the message type split between the different classes where Class A vessels are international voyaging ships greater than 300 gross tonnes and all passenger vessels (regardless of size), and Class B vessels are for smaller vessels; including commercial, fishing, recreation and leisure vessels[3].

	Class A	**Class B**
Ship Type	>300GT Legal Requirement	Small vessels Not compulsory
Dynamic	1,2,3	18
Static	5	19
Dynamic and Static	NA	24

Table 2.2: AIS message type class breakdown.

The message types; 1, 2, 3 and 18 are referred to as dynamic messages, while message types; 5, 19, 24 contain static information. Addendum here: Type 5 messages contain both static and semi static information (will be covered more in a later section).

2.7.2 Message Structure

```
!AIVDM,1,1,,B,13P:VTP000wjAKpNTcnP502R05‘‘,0*6F
!AIVDM,1,1,,A,33M@F30P@uOj:g8NUa1UOJ2>01t0,0*19
```

Figure 2.18: Example of AIS Message Types 1 and 3

2.7.3 Payload Content

Over the 27 message types there are over 100 different variables used. And each type has a subset of these and the MMSI is the primary key across all types, these include:

- Vessel name
- Vessel position
- Vessel speed over ground
- Vessel course over ground

[3]Class B is completely voluntary and the AIS equipment is not as powerful as class A AIS equipment therefore, the range a class B AIS message is significantly shorter than that of a class A message.

- MMSI
- IMO Registration number
- Ship Type
- Radio Call Sign
- Length
- Beam - this is the width of a vessel
- Draught - the depth of the vessel below the waterline. Note this changes journey to journey as it depends on the mass of the cargo being carried.

2.7.4 The MMSI Number

The Maritime Mobile Service Identity (MMSI) is a unique number that is assigned to the vessel and all AIS messages will include this unique identifier. The structure of a MMSI number is made up of 9 digits (note it is possible for there to be 10 (e.g. Man overboard, Search and Rescue) - but some systems can only handle 9 thus a 9-digit MMSI 123456789 when prefixed with a "9" becomes the 10-digit 9123456789 and the 9-digit system will cut this down to the 9-digit MMSI 912345678).

The MMSI number is the unique identifier (think vehicle registration plate) for the vessel and the number will only change if there is a change of ownership of the vessel (if a nation recycles their MMSI numbers this may change on a month by month basis). The first 3 digits are the Maritime Identification Digits (MID) [60] which denote the vessel's flag state.

2.7.5 The IMO Registration Number

Reference [123] made it mandatory for [52] to regulate the use of IMO registration numbers as ships identification. This 7-digit IMO registration number is a requirement for all vessels over 100 gross tonnage to be identifiable by an IMO registration number (with a few exceptions). The IMO registration is assigned to the hull of a vessel and will remain with that hull for the lifetime of the vessel despite ownership (think Vehicle Identification Number).

The IMO registration number can be found in the static AIS messages such as the type 5 messages. Due to the nature of the IMO registration number there is a method for ensuring that an IMO registration number is valid. The validity is verified by the check digit (7th digit of the IMO registration number).

The check corresponds to a multiplication sum over the first 6 digits of the IMO where the units digit of the final sum must match the check digit (7th digit of the IMO). For example, validating the IMO registration number 9652806;

$$9 \times 7 + 6 \times 6 + 5 \times 5 + 2 \times 4 + 8 \times 3 + 0 \times 2 = 15\mathbf{6} \quad | \quad 9652806 \tag{2.53}$$

Since the right most digit of the sum (6) equals the right most number of the IMO number (6), this IMO registration number is valid.

The following are not valid IMO registration numbers.

1111111;

$$1 \times 7 + 1 \times 6 + 1 \times 5 + 1 \times 4 + 1 \times 3 + 1 \times 2 = 2\mathbf{7} \quad | \quad 1111111 \tag{2.54}$$

1231231;

$$1 \times 7 + 2 \times 6 + 3 \times 5 + 1 \times 4 + 2 \times 3 + 3 \times 2 = 5\mathbf{0} \quad | \quad 1231231 \tag{2.55}$$

2.7.6 AIS Ship Type

The ship type is recorded as a double digit in the range 10-99. The first digit denotes the primary ship type and the second digit provides additional information, either about the class of a vessel in a given ship type, if a vessel is carrying hazardous or if a vessel is performing Search and Rescue. Table B.5 provides the full list of ship types in the range 10-99.

2.8 Datasets used in this study

As already discussed this study used AIS data augmented with the IHS Markit World Register of Shipping database.

Three datasets of AIS data have been sourced. Each one of the datasets provides a different subset of various AIS messages. The properties of the datasets can be seen in Table 2.3. There are datasets that contain AIS reporting at the native rate which is the rate at which the AIS base stations are receiving messages (this can be seen in the North Atlantic dataset in Figure 2.23 and the Merseyside dataset in Figure 2.19) and in hourly snapshots which can be seen in the global dataset (Figure 2.29).

	Local	Regional	Global
Area	Merseyside	North Atlantic	Global
Area of coverage	$2,945\text{km}^2$ (0.0006%)	$44 \times 10^6\text{km}^2$ (8.6%)	$510 \times 10^6\text{km}^2$ (100%)
Provided by	Denbridge Marine	Exact Earth	IHS Markit
Duration	1 Day	32 Days	15 Days
Start	2018-05-29 09:25:02	2017-08-10 00:00:00	2017-01-23 00:31:00
End	2018-05-30 10:35:09	2017-09-10 20:59:58	2017-02-06 23:31:00
Quantity	200,000	68,000,000	18,100,000
Frequency	2-10 Seconds (Native Rate)	2-10 Seconds (Native Rate)	Hourly
Message Types	Fused Dynamic and Static	All 27 Message Types	Fused Dynamic and Static
Unique MMSIs	116	115,598	62,103

Table 2.3: Properties of the datasets.

AIS data is available for live coverage [49, 88] and historic coverage [37, 49]. Additionally, Merseyside-based Denbridge Marine[4] provide AIS data from a local AIS base station situated on the Northern most point of the Wirral peninsular at Fort Perch Rock.

[4] An ICASE industry partner of another PhD project within the University of Liverpool research group.

2.8.1 Merseyside Dataset

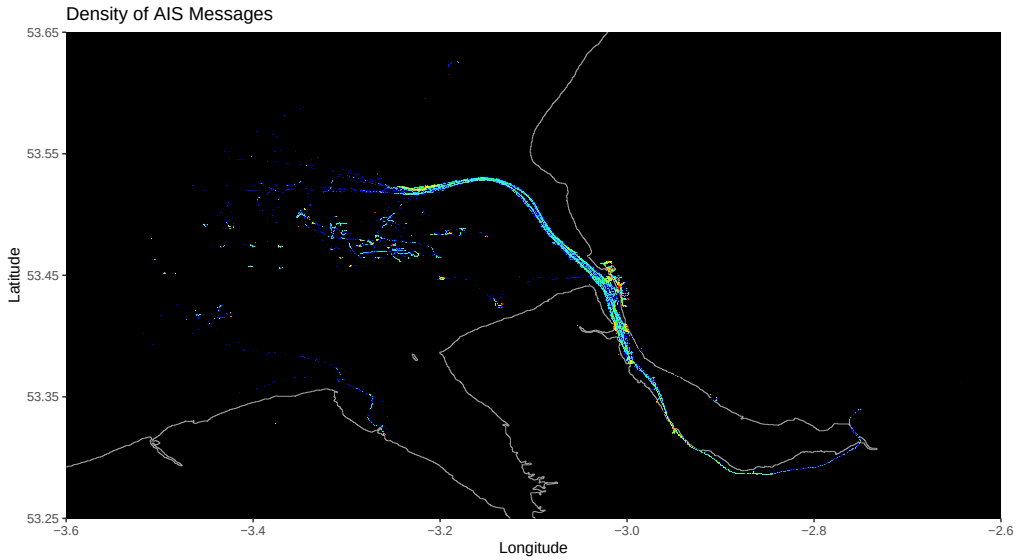


Figure 2.19: A 24 hour period from a single AIS receiver at Fort Perch Rock located at the Northernmost tip of the Wirral Peninsula. Density resolution at 2 arcseconds. *(Data provided by Denbridge Marine)*

Figure 2.19 depicts the density of AIS observations for the Merseyside dataset. Of the 116 MMSIs in the dataset, 0.86% have an invalid MID. Figure 2.22 shows the distribution of MID flag states. There is a 1-to-1 match between MMSIs and IMOs (for a given MMSI, there is only one IMO and for a given IMO, there is only one MMSI). The IMO Checksum can be used to validate the IMO number. All IMOs are validated. 47.36 % of vessels report a length of 0m and a beam of 0m. There are no $(91°N, 181°E)$ messages in the dataset. Figure 2.20 presents the difference in distance and in time for all measurements in the Merseyside dataset and includes the confidence interval for the maximum speed of a vessel extracted from the WRS database. Figure 2.21 presents the elapsed time between measurements, here depicting that majority of the data us an update rate less than 60 seconds.

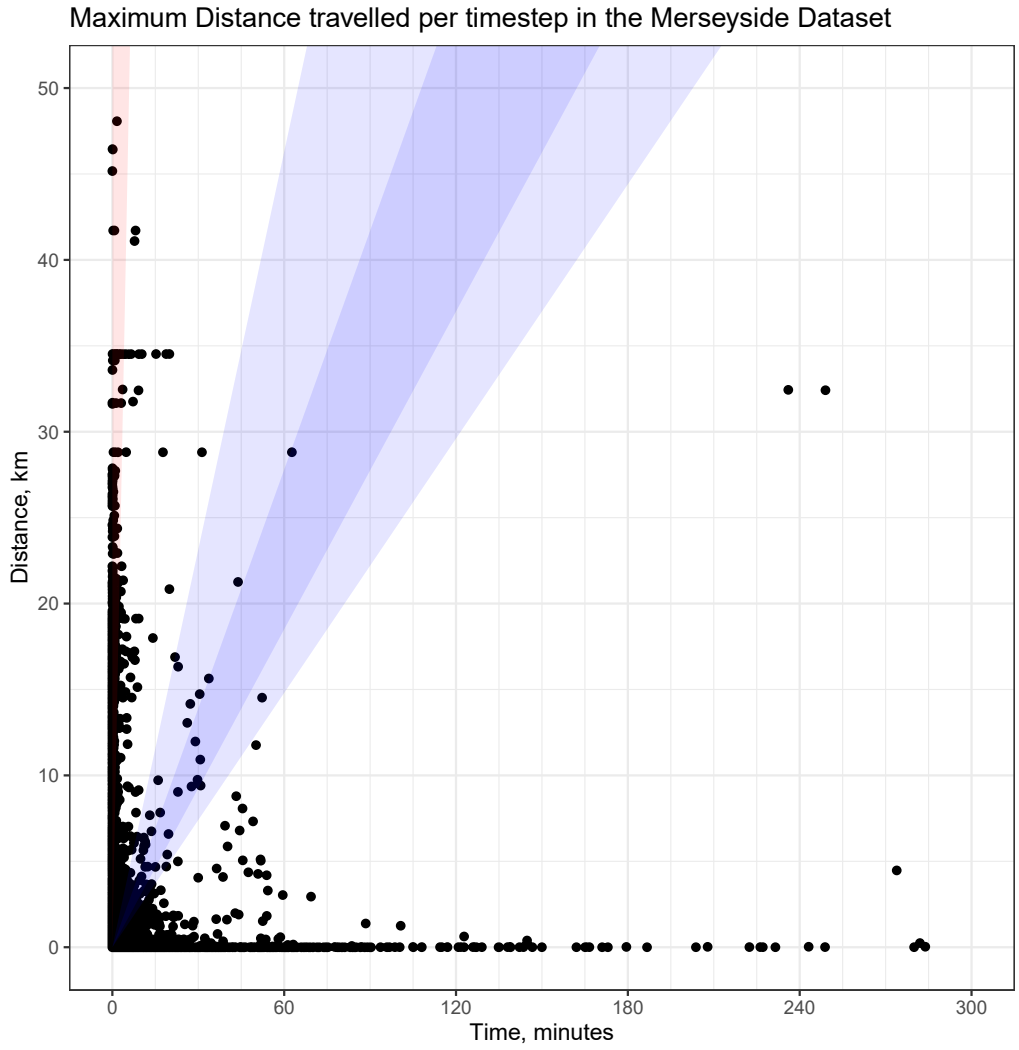


Figure 2.20: The figure represents the maximum distance a vessel travels over a given time from the Merseyside dataset. The blue depicts the 95% confidence interval and the dark blue depicts the 68% interval, where 68% of all ships are in the dark blue and 95% are in the light blue derived from the vessel maximum speed in the WRS database. The red region depicts vessels travelling faster than the water speed record of (511kmph).

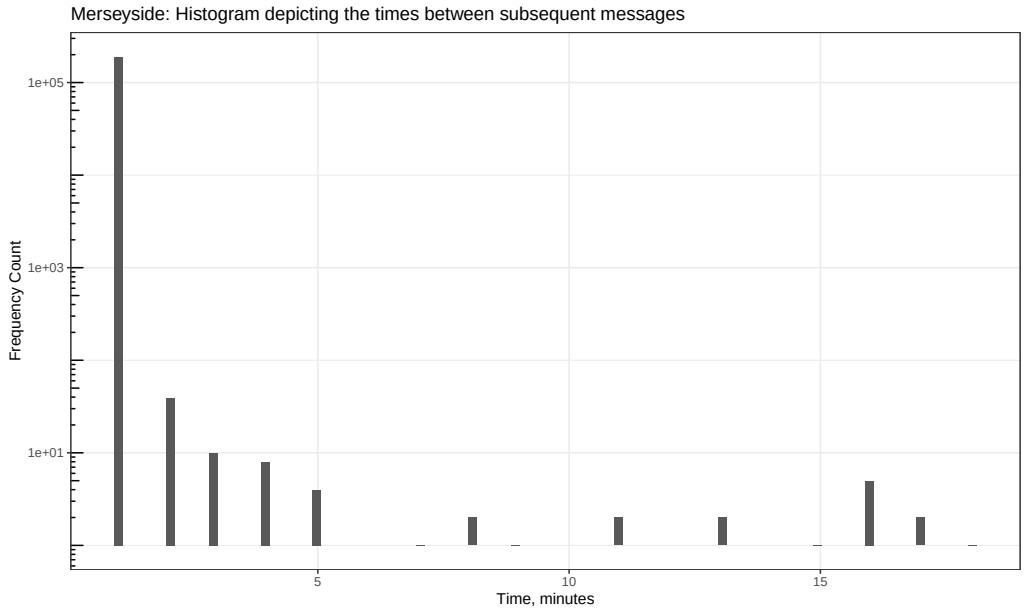


Figure 2.21: Elapsed time between subsequent AIS messages of the same MMSI.

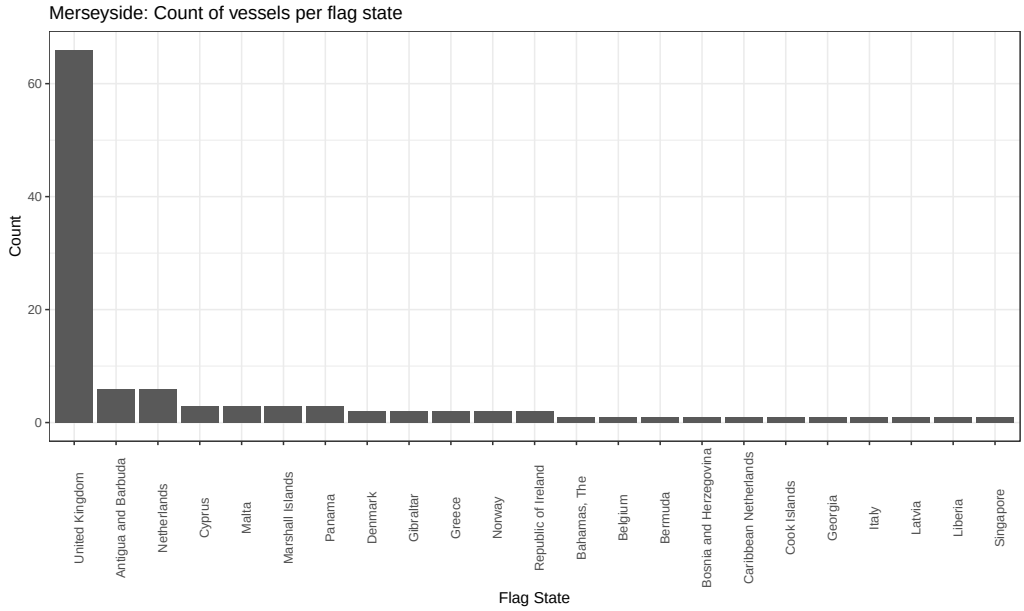


Figure 2.22: Count of MMSIs with valid MID number per flag state for the Merseyside dataset.

2.8.2 North Atlantic Dataset

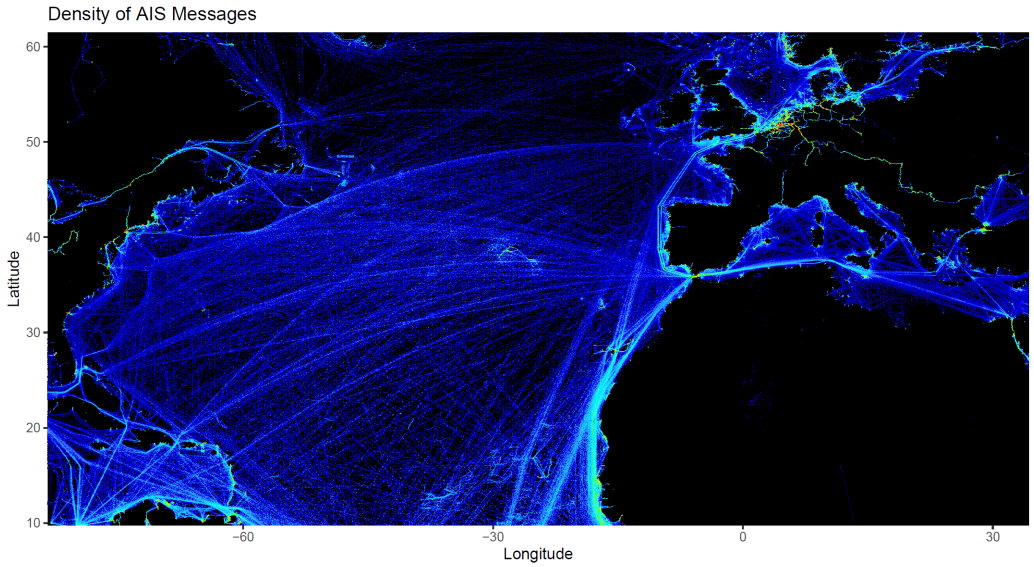


Figure 2.23: A 32 day period from an aggregated commercial AIS data source for the North Atlantic. Density resolution at 3 arcminutes. *(Data provided by Exact Earth)*

Figure 2.23 depicts the density of AIS observations for the North Atlantic dataset. Of the 115,000 MMSIs in the dataset, 1.88% have an invalid MID (2,179 MIDs are invalid). Figure 2.24 shows the distribution of MID flag states for the top 23 countries. Figures 2.25 and 2.26 show that there are multiple IMOs for a given MMSI (there is more than one vessel using a single MMSI) and there are multiple MMSIs for a given IMO (a vessel is using more than one MMSI). Since this dataset provides the dynamic messages and static messages separately, there are 31% of MMSIs that have orphan dynamic messages (i.e., there is no corresponding static message for the given MMSI). All MMSIs from static reports are accounted for in the dynamic messages. The IMO Checksum is used to validate the IMO number and 93% of IMOs are valid (1,736 IMOs are not valid). 2.91 % of reports have a length of 0m which corresponds to 4.74% of MMSIs. 3.03% of reports have a beam of 0m which corresponds to 4.92% of MMSIs. Figures 2.27 and 2.28 show frequency of the number of lengths and beams reported by a MMSI respectively. 11.34% of static messages report a draught of 0m which corresponds to 11.73% of MMSIs. 11.83% of reports have a draught greater than their length which corresponds to 7.16% of MMSIs. There are 0.32% of $(91°N, 181°E)$ messages in the dataset.

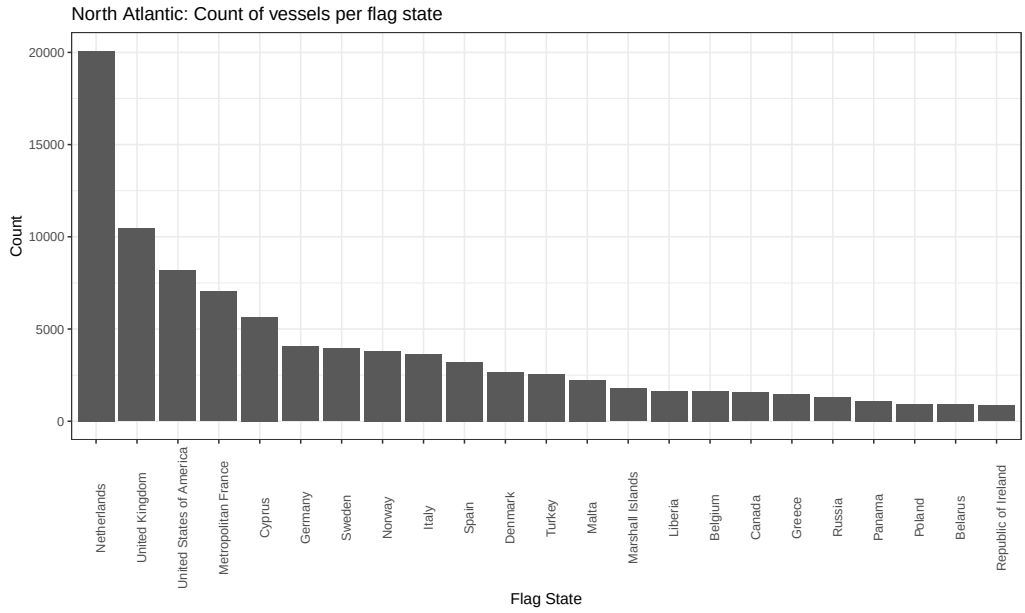


Figure 2.24: Count of MMSIs with valid MID number per flag state for the 23 most common countries for the North Atlantic dataset.

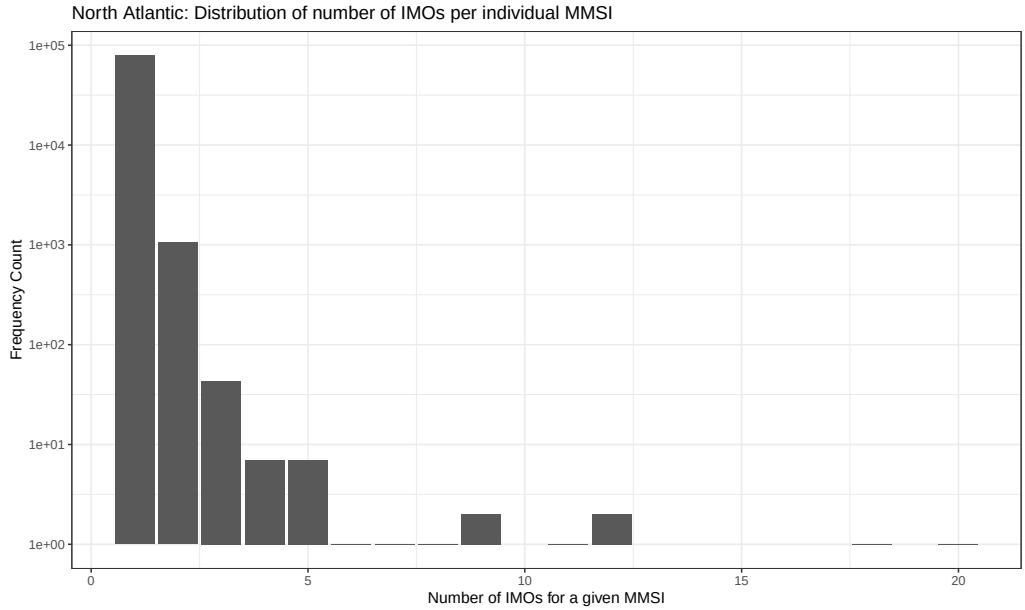


Figure 2.25: Count of the number of IMOs per MMSI for the North Atlantic dataset.

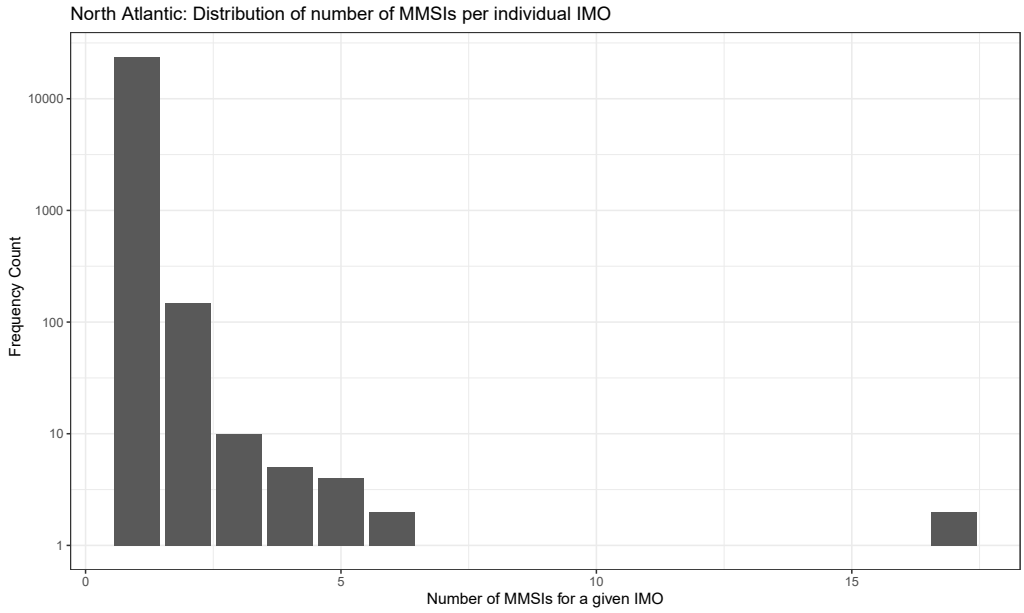


Figure 2.26: Count of the number of MMSIs per IMO for the North Atlantic dataset.

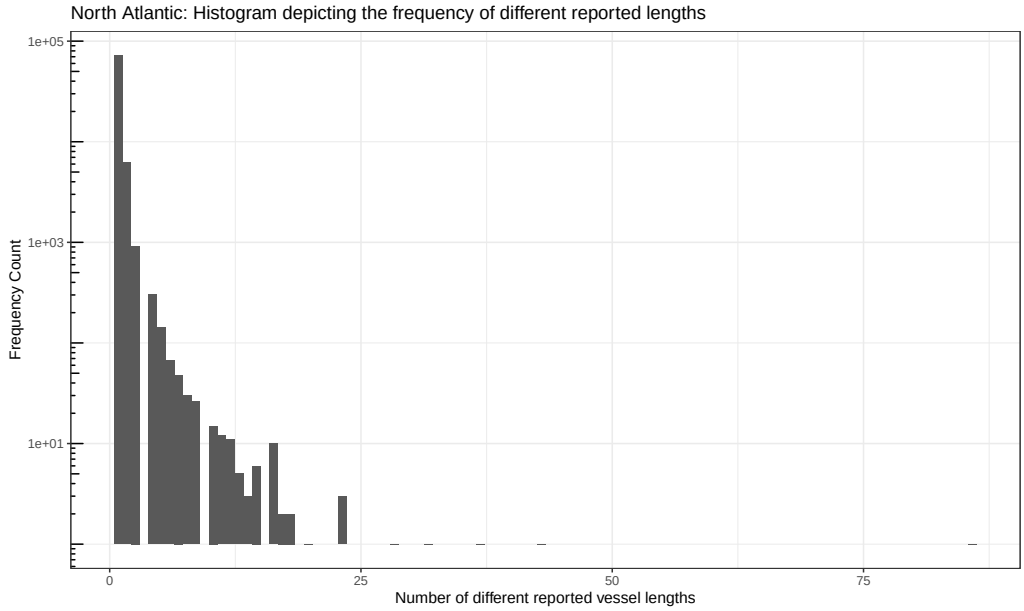


Figure 2.27: Count of number of different lengths per MMSI reported in the static reports for the North Atlantic dataset.

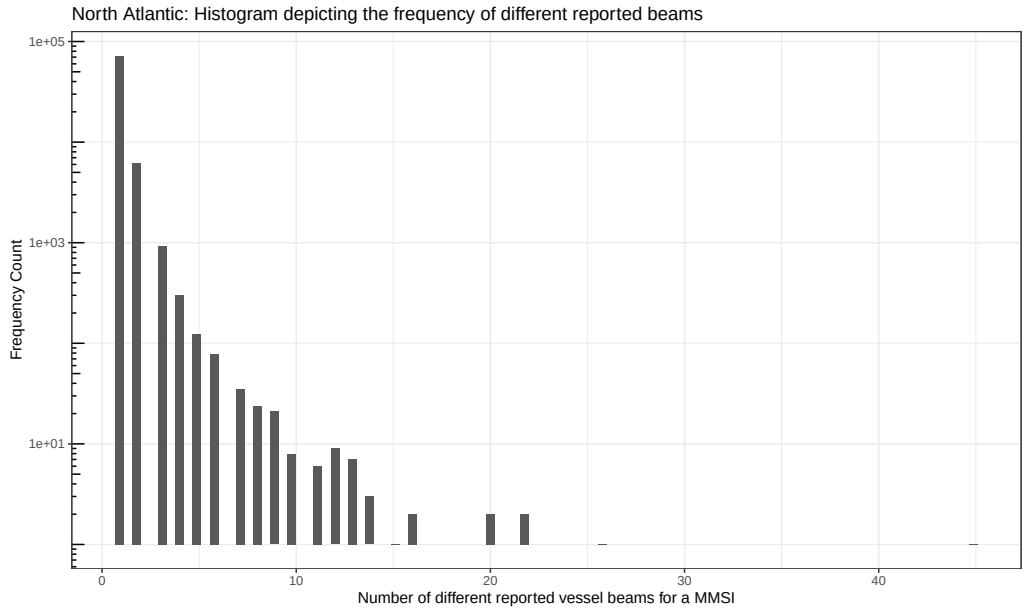


Figure 2.28: Count of number of different beams per MMSI reported in the static reports for the North Atlantic dataset.

2.8.3 Global Dataset

A sample AIS dataset was acquired from AISLive and recorded by IHS Markit. The data consisted of 18,197,202 AIS reports recorded between 00:32 23rd January 2017 and 23:31 6th February 2017, in a worldwide geographical area. This provided 15 days of hourly observations. The dataset consists of approximately 147,000 vessels (see Figure 2.29).

From figure 2.30, there are MMSIs with observation counts greater than the maximum possible for the fixed time steps in the IHS data. These MMSIs (a subset of which are visualised in figure 3.13) depict that there are more than one vessel reporting on each of these MMSIs. On further investigation of the IHS dataset, this can also be seen in those MMSIs with observations less than 360. These vessels are only reporting for a small interval within the 15 days of the dataset.

The AIS specification states the MMSI number is made up of 9 digits. The number of AIS reports in this dataset with 9-digit MMSI numbers is 99.9% where 20,839 display other numbers of digits ($\in \{1, \ldots, 10\} \backslash \{9\}$). The details of inaccurate length and beam being reported include;

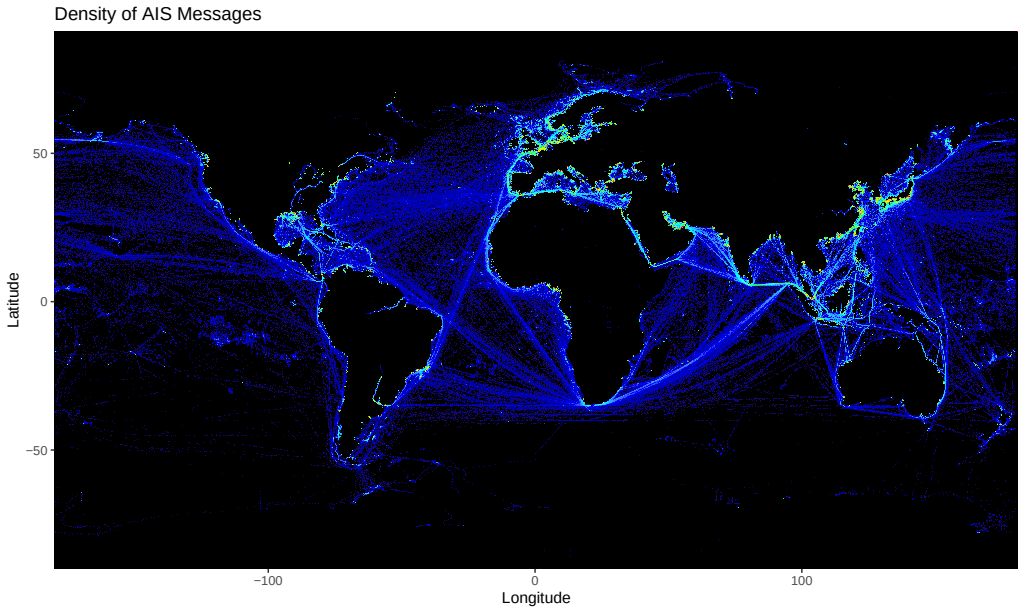


Figure 2.29: A 15 day period from an aggregated commercial AIS data source for the world where the AIS reports are hourly. Density resolution at 10 arcminutes. *(Data provided by IHS Markit)*

- 3.9% of AIS reports had length 0m.
- 3.3% of vessels reported 0m for their length (identifier *key*).
- 4.3% reported different lengths when using MMSI as the identifier (0% when using *key*).
- 3.4% of vessels reported a beam of 0m.

Figures 2.32 and 2.33 show that there are multiple IMOs for a given MMSI (there is more than one vessel using a single MMSI) and there are multiple MMSIs for a given IMO (a vessel is using more than one MMSI). All 59,511 IMO numbers are valid IMO numbers. 5.2% of AIS reports had draught of 0m and 3.3% of vessels reported a draught of 0m. It was also noted that 3.4% of vessels had a draught greater than that of the length of the vessel. The data consists of 149,100 unique recorded destinations. 5% of the AIS observations provided no destination variable. Approximately 22% had destination information that was easily to be understood. 16% provided non-informative information which predominately

consisted of a variety of punctuation. A large portion of the recorded destinations provide difficult to interpret abbreviated destinations. 0.6% of observations are ($91°N$, $181°E$) error messages. Figure 2.31 show the distances between consecutive measurements is extremely high. This dataset has an hourly update rate which suggests there are either more than one vessel reporting on a MMSI or the vessel can travel over 10,000kmph.

Identifiers	Ship Details	Movement
IMO Number*‡	Ship Type*	Latitude†
MMSI†‡	Beam*	Longitude†
Call Sign*	Draught*	Speed†
Ship Name*	Length*	Heading†
	Additional Information*	ETA*
		Destination*
		Movement Date & Time*
		Movement ID*
		Move Status*
		Time†

Table 2.4: Features of the global AIS dataset(provided by IHS). Due to this data being pre-fused by IHS Market, these are a combined set of variables that are normally across multiple AIS message types (see Table B.1. Items denoted with a dagger (†) text denotes the fields found in dynamic messages such as types 1, 2, and 3. Items marked with an asterisk (*) denote semi-static and static information predominantly found in type 5 messages. The items marked with a double dagger (‡) refer to the primary identifiers used by AIS (MMSI), and WRS (IMO).

	MMSI	IMO Number	Call Sign	Description
>360	216	739	535	The number of objects with observations over 360
MAX	913	722	66,395	The largest number of observations for a single object
#	62,103	59,511	60,737	The number of unique objects for given identifier

Table 2.5: Unique Identifier Analysis

Table 2.5 shows that of the set of identifiable variables in this dataset, none of them are actually unique. The dataset consists of 360 total observations, this means if a variable was to be unique, there should not be any identifier having more than 360 observations.

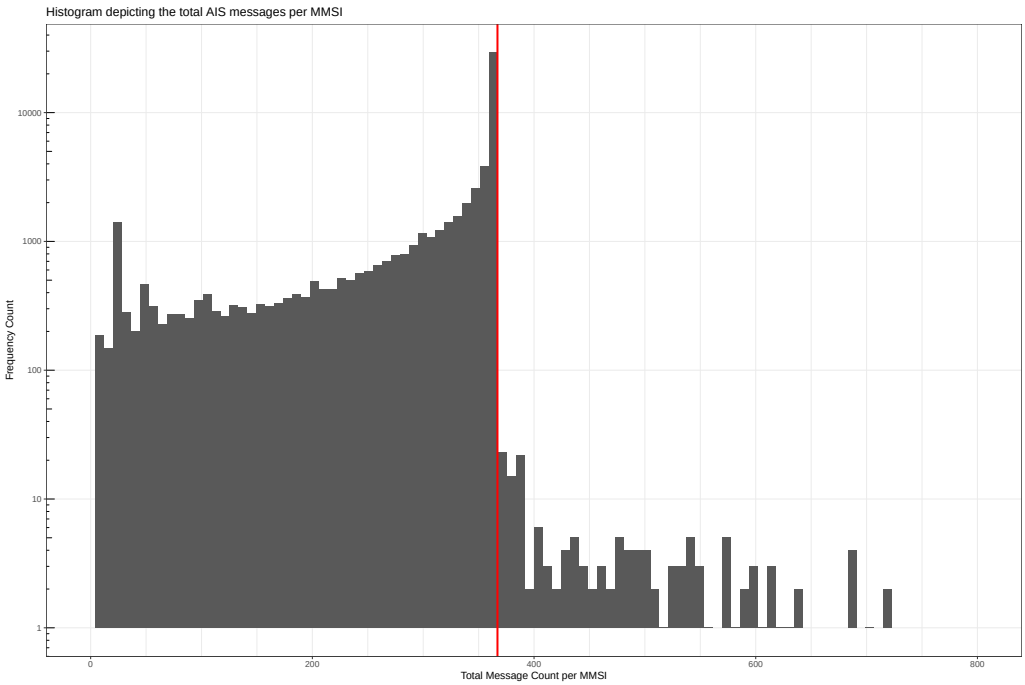


Figure 2.30: The IHS Dataset is has hourly reporting rates which means in the 15 day dataset a vessel can only have reported 360 times. The vertical red line denotes the split between those quantity of vessels with their total message count below this threshold and the quantity of vessels with more than 360 reports.

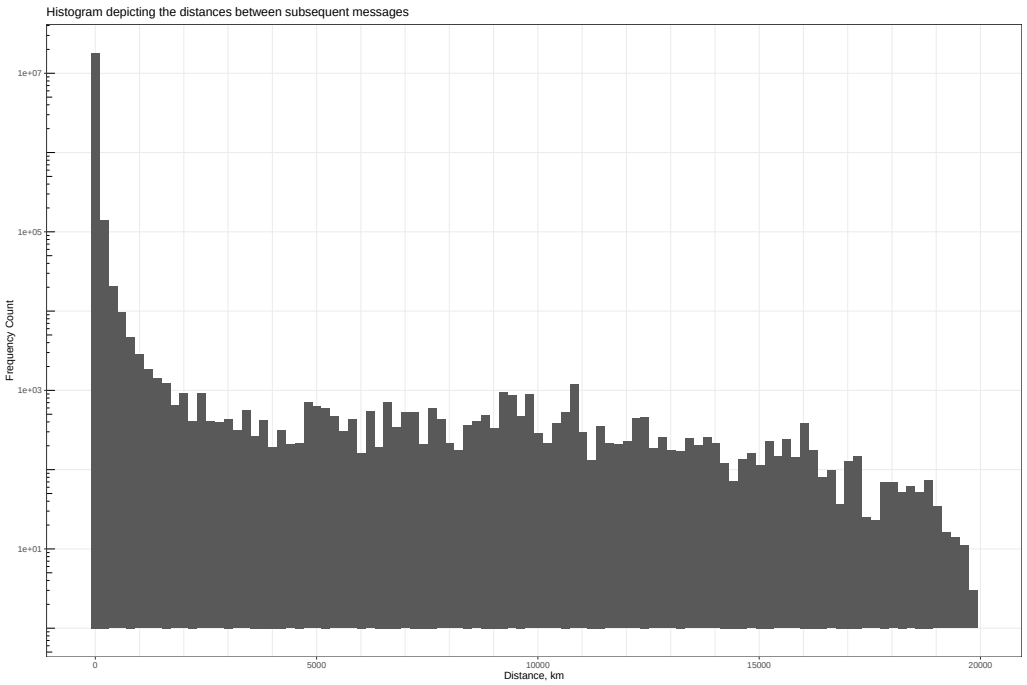


Figure 2.31: Frequency of distances between consecutive observations.

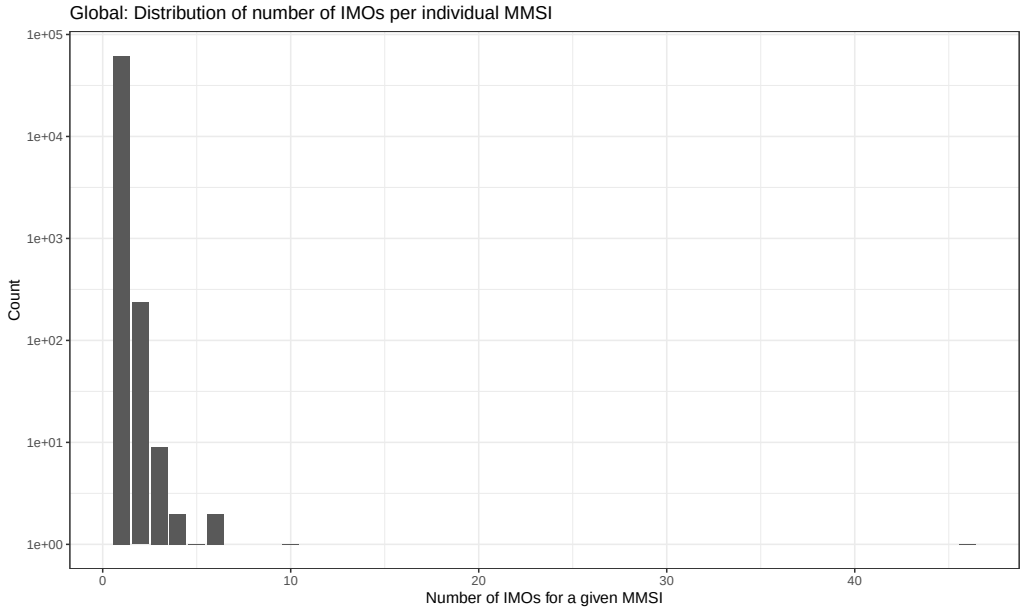


Figure 2.32: For a given MMSI, the count here is for the number of multiple IMOs.

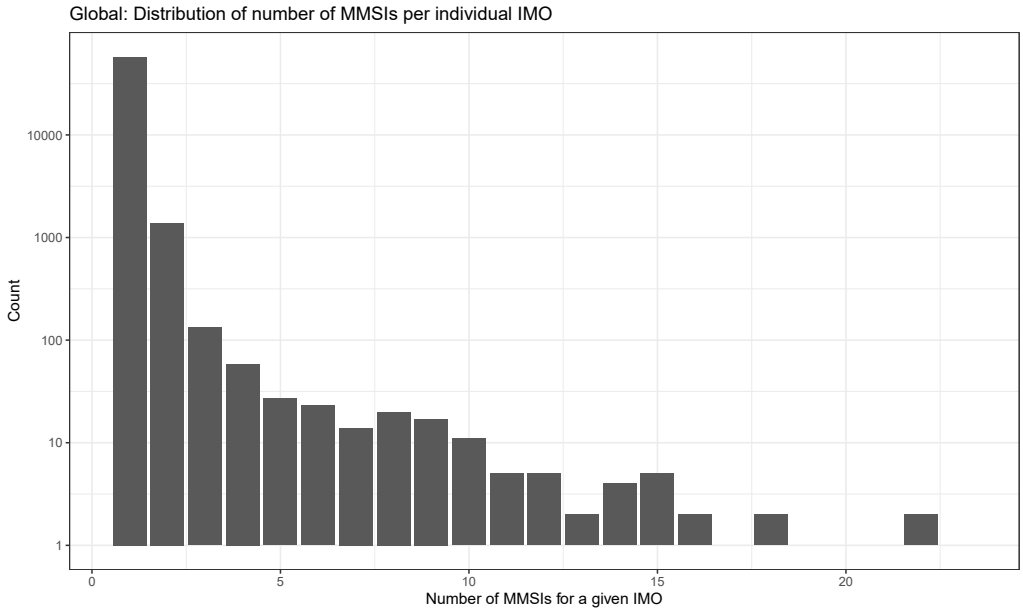


Figure 2.33: For a given IMO, the count here is for the number of multiple MMSIs.

2.8.4 Peculiarities

The challenge is to determine which of the targets are maritime vessels and which are land based. Additionally, which targets are broadcasting either the wrong identity or another vessel's identity. A large problem arises when multiple vessels use the same MMSI number. While it is difficult to track historical data of a vessel which uses a duplicative MMSI, if another feature is correct, the vessel's identity can be verified. This concatenation of AIS features provides a successful identifier, there still are outliers in the modified dataset. Spatial analysis is utilised to add an additional level of data fusion validation. The MMSI is the unifying key value across all AIS message types which means that if more than one vessel is reporting on a MMSI the static messages are unable to be assigned to the correct positional messages.

The issues arising are that it is difficult to track a unique target over time as there is no fixed (unique) variable that can be used for identifying vessels across message types[5]. There is no single identifier (IMO, MMSI, Callsign, etc. See Table 2.4) that meets the criteria for every vessel to be uniquely tracked.

Over the course of several days of discussions, it was clear that the global dataset (see Section 2.8.3) was not an adequate dataset for comparison with the system currently implemented at the NMIC. This led to the acquisition of the North Atlantic dataset from Exact Earth [37, 13, 97] and the Merseyside dataset provided by Denbridge Marine [33] which provided comparative data environments to the data used by the NMIC. As well as understanding the AIS system and associated data, it was important to get an understanding of the true behaviours of vessels behind transmitting their messages via AIS. A set of behaviours has been defined that covers behaviours of interest to the NMIC (See Figure 2.34).

A: Normal vessel behaviours

B: Normal vessel behaviours with small number of measurements with latitude/longitude shifts.

C: Normal vessel behaviour with a large amount of measurements with latitude/longitude shifts.

D: Normal vessel behaviour but the MMSI of the vessel changes.

[5]The IMO number is unique to a vessel's hull and remains with it for the life of the vessel. A MMSI number is issued by the flag state and might change due to MMSI recycling by the flag state, sale by the owner to a company in another flag state over the course of a vessel's life span.

E: Normal vessel behaviour but a message has been mis associated with another MMSI number.

F: Two vessels with normal behaviour reporting on the same MMSI

G: Two vessels with normal behaviour reporting on the same MMSI with small number of measurements mis associated.

H: Two vessels with normal behaviour reporting on the same MMSI where the duplicate vessel begins to mimic the true vessel's MMSI.

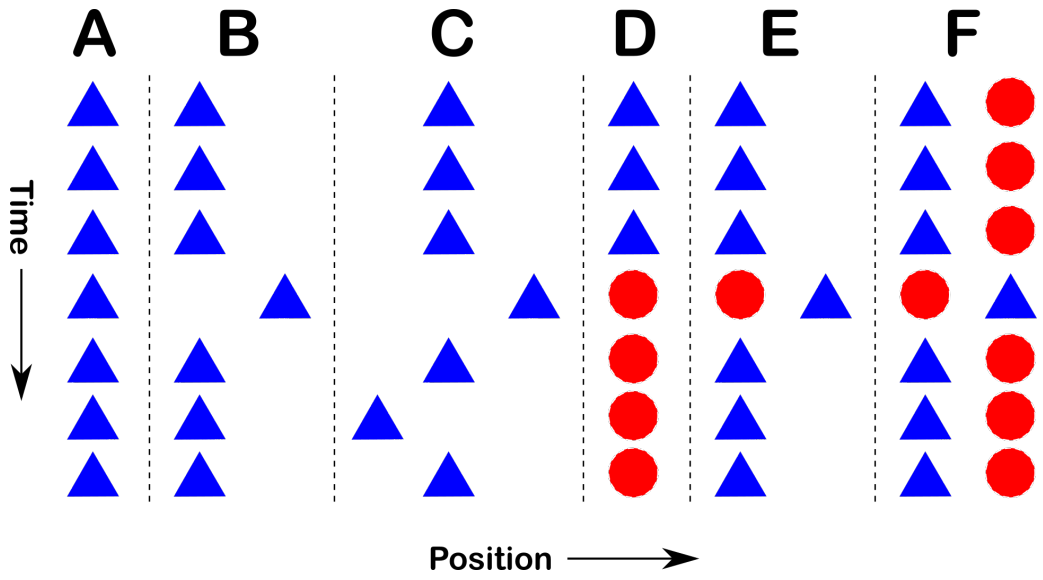


Figure 2.34: MMSI Exemplar

2.9 Analysis of Error Messages

This section provides an overview of the types of errors in the datasets described in Section 2.8 and looks to understand the prevalence of GPS error messages and periods of time that have no reports.

There are many causes of reported position errors received in AIS messages. These include;

- No GPS. Causes could include; the GPS aerial being disconnected or even hardware failure
- Bad reception of GPS signal leading to errors in GPS lock. This is the main reason for receiving crazy tracks.
- GPS Jamming.
- Congestion in the VHF band leading to bit errors. The AIS checksum is added by the receive station and is not a check of the originating transmission - a fundamental failure of the AIS standard
- Local geography / reflections / multipath GPS propagation
- Hardware failure
- Sticky bits in the transmitter DAC etc.
- Sporadic errors
- Full blown spoofing of a vessels position, sometimes these manifest as a shift in latitude or longitude.

We also have challenges with missing data, in some cases this could be because the vessel has turned its AIS transponder off, however this would be unusual. In most cases the reasons for missing data would be because the vessel is out of range of a coastal AIS receiver and there are no overhead satellites about to receive the signal. This can be compounded further when comparing detection ranges for class A vessels with class B vessels. Class B vessels broadcast their positions with a lower transmitter power and hence their signals don't propagate as far as for class A vessels.

If a vessel is no longer being detected either by terrestrial base stations or satellites, and we assume it is in range, then the AIS transmitter might be at fault. The fault of the AIS transmitter may be down to user error, faulty equipment, or a deliberate act. We refer to this case as "going dark" or the vessel being a "dark target". Using AIS data alone it is impossible to determine why a vessel went dark. Clues as to why can be gained from information such as the position of the last transmission before going dark and the position of the first message back. We can use these positions to judge the probability that the vessel has gone dark deliberately which would be high in areas of good coverage and

low in other areas. We could also use such data to obtain profiles of places where vessels potentially deliberately go dark.

Fortunately, most AIS messages received are correct. However, there are also a number of mechanisms where the information (particularly location) could be incorrect. There could be transmission errors where the message gets damaged in transit. These errors will be referred to as bit errors. If the bit error in the vessel position, the latitude and/or the longitude could suddenly add jump by a large amount (i.e., an ocean!). If the bit error is in the sign of the latitude or longitude, the position will be flipped in the equator or prime meridian. If the bit error is in the MMSI, the result will be that a new vessel will be assigned the position report. The changes could be deliberate. There are cases and potential scenarios that include vessels, augmenting their positions by a "bit shift".

There are occurrences when an AIS message provides a GPS error (91 degrees North and 181 degrees East) this is the equivalent of a `NULL`. These messages are caused by a lack of GPS signal.

This section shows the prevalence of such errors that need to be taken into account when constructing a picture of vessel movement.

2.9.1 Analysing the Tracklets for Errors

An analysis of the output of the multiple target tracker was carried out to try and determine an estimate of the number of vessels sharing a MMSI value (there should only be one) and to assess the variation of error messages per vessel.

Position reports of the form ($91°N$, $181°E$) are treated as a position error. Figure 2.35 shows the error reporting percentage per MMSI of the global dataset. The provider of the Merseyside dataset removed error messages in their cleaning process and as a result we cannot produce the same analysis on the Merseyside dataset.

Vessels that have gone "dark" have ceased to transmit their AIS messages. These dark vessels are typically detected when fusing the cooperative AIS data with a non-cooperative data source such as radar that can provide you with the position of a vessel that has no associated AIS messages. Figures 2.36, 2.37 and 2.38 show the times between consecutive measurements.

We are able to use the understandings that a vessel should be reporting regularly, and there are areas with no AIS receiver coverage, to calculate the expected time between tracklets. (Note: If the distance and time were small enough the position reports would

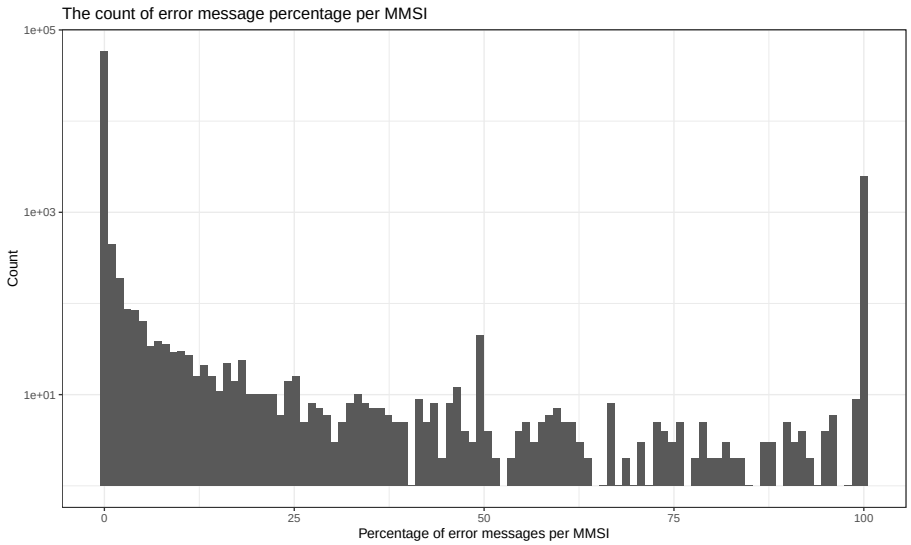


Figure 2.35: The count of percentage error per MMSI. This graph shows that for the majority of vessels position reports are received with no errors however there are more than 1000 vessels that report a Null position report.

have been tracked and incorporated within a tracklet. Also, if the time between points is just outside of the disambiguation tracker's thresholds, the automatic track stitching could still associate the reports.) The following figures show the results over the three datasets. Figures 2.39, 2.40 and 2.41 illustrate the time interval between tracklets where no measurements were observed for each dataset. Figures 2.42 and 2.43 provide a zoomed in representation of Figure 2.41 over the Mediterranean (Figure 2.42) and the South China Seas (Figure 2.43).

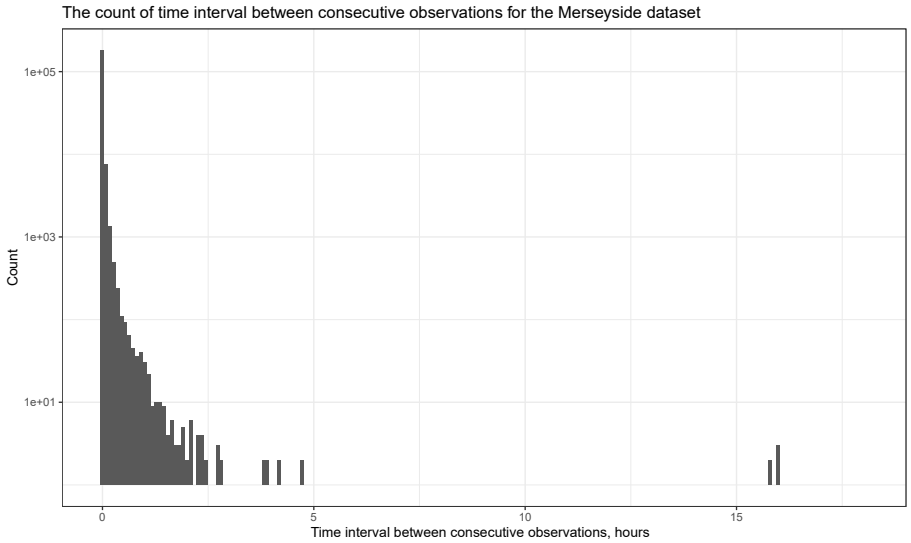


Figure 2.36: The count of time interval between consecutive observations for the Merseyside dataset. The majority of observation update rates are as expected less than 5 minutes.

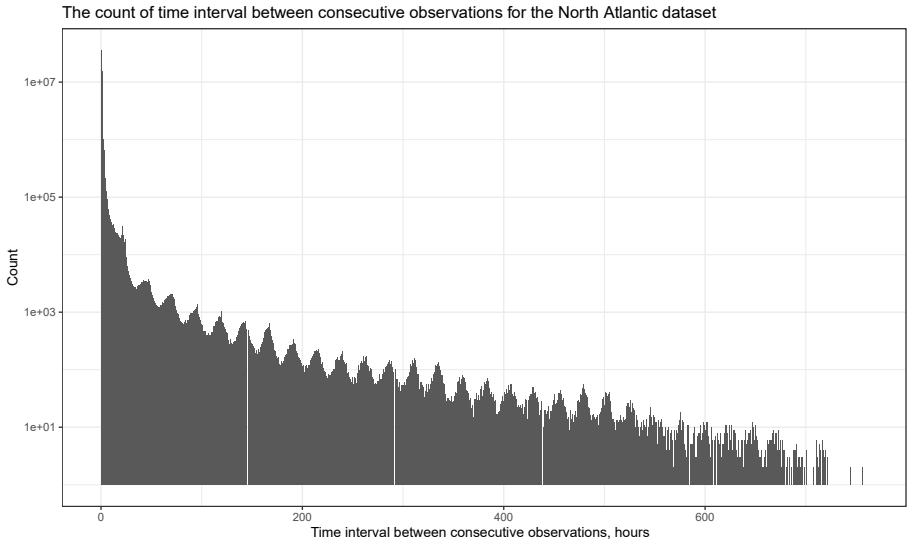


Figure 2.37: The count of time interval between consecutive observations for the North Atlantic dataset.

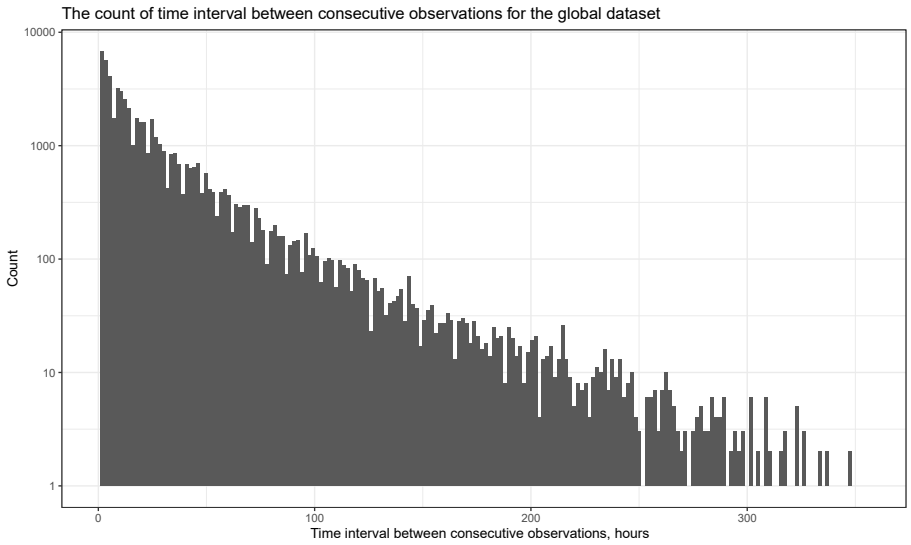


Figure 2.38: The count of time interval between consecutive observations for the global dataset.

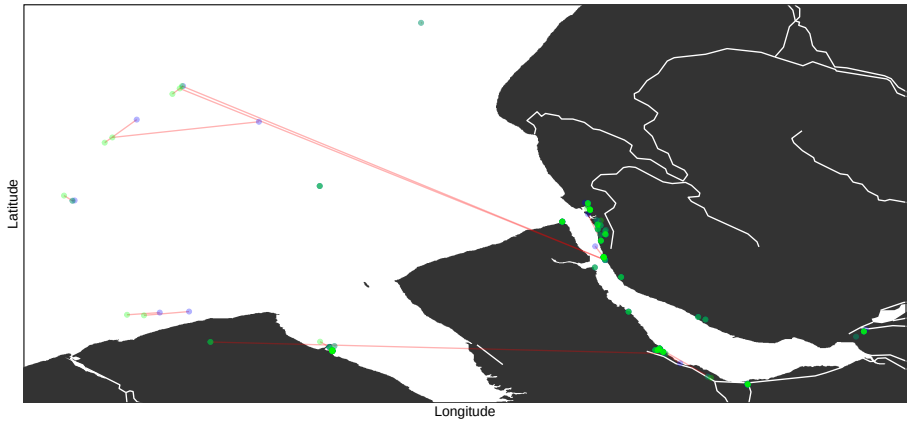


Figure 2.39: Map showing Merseyside dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 1 hour.

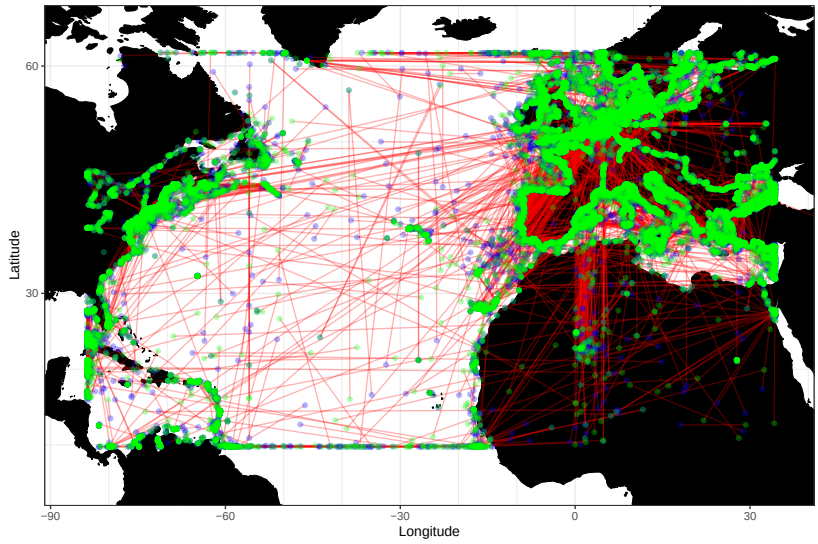


Figure 2.40: Map showing the North Atlantic dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is > 5 days.

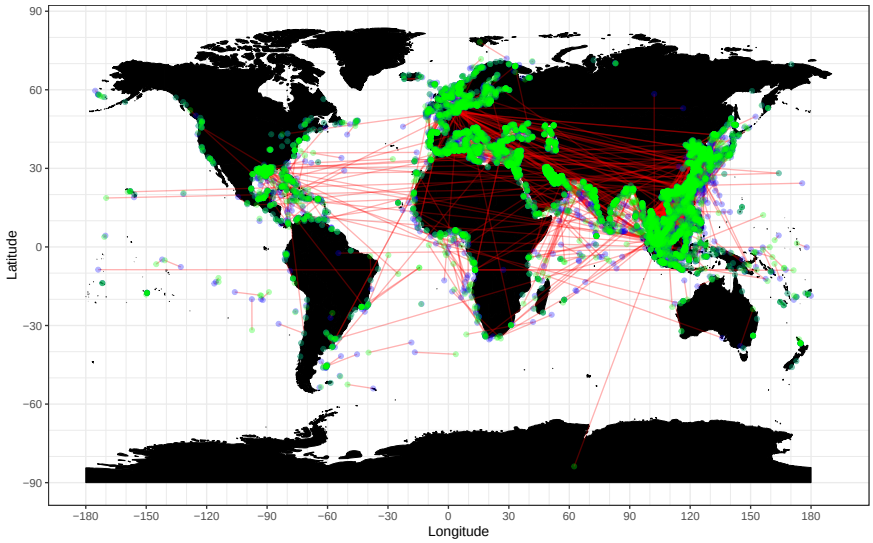


Figure 2.41: Map showing the global dataset. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours.

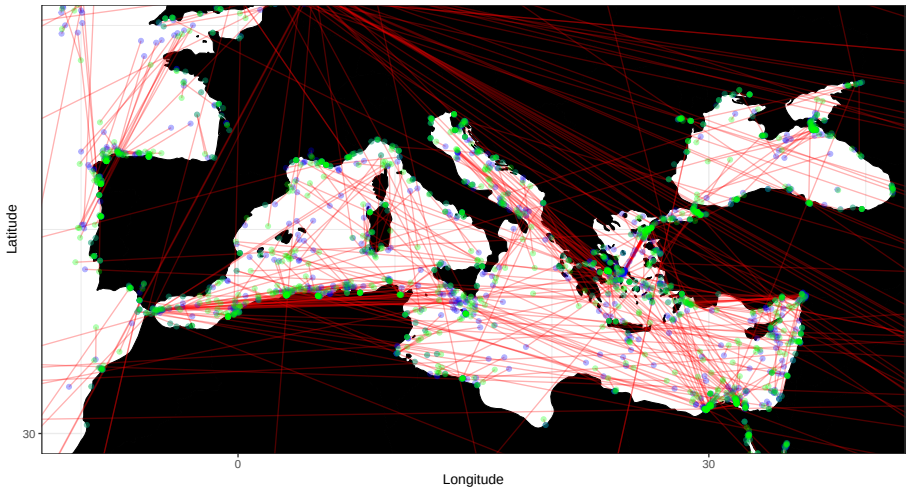


Figure 2.42: Map showing the Mediterranean extracted from the global data. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours.

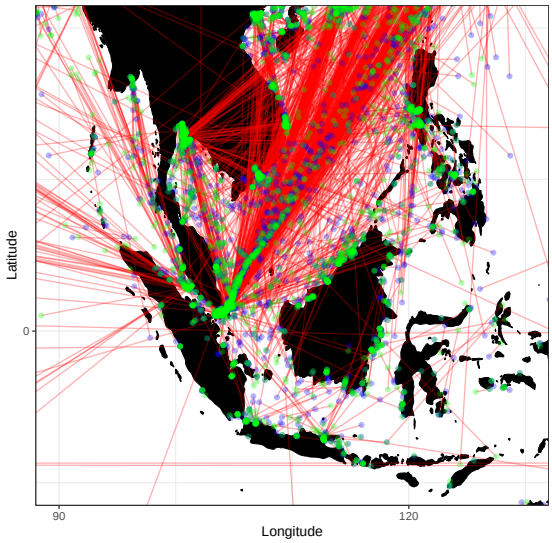


Figure 2.43: Map showing the South China Seas extracted from the global data. The red lines join reports of the same vessel where there are large gaps between received signals. In this example the time interval is $>$ 48 hours.

2.9.2 Discussion of Findings

These maps illustrate that there are four scenarios when describing 'dark' targets. These can be described as:

- Vessels that are in the open ocean and out of range of AIS receivers
- Vessels that travel from open ocean to coastal areas
- Vessels that travel from coastal areas to open-ocean
- vessels that stay near to the coast and within AIS coverage

As AIS reception is more sporadic in open ocean it is more likely to get vessels that appear to be going dark due to lack of coverage. It is much more likely that vessels in coastal areas will be detected, and hence missing position reports could indicate a vessel trying to hide.

During this analysis it was hoped that different behaviours, as shown above, would manifest in different distinct clusters, such as an open-ocean cluster and a coastal cluster, however, these cases seem to be truly random.

2.10 Summary

This chapter has provided an introduction to the mathematical concepts that are used in the following chapters. The state estimation and tracking discussed in Section 2.1 is implemented in Chapter 3 and the track stitching described in Section 2.2 is used in Section 4.1. The text analytic methods discussed in Section 2.3 is applied to the abstracted data in Section 5.1 and the change point detection described in Section 2.4 is applied to the dataset grouped into regional counts in Section 5.2.

Chapter 3

Disambiguation

This chapter describes the problem space in the context of multiple target tracking and concludes with the comparison of the raw data to the disambiguated data.

The following chapters follow the same structure; providing an overview of the problem and its chapter's solution, specific literature providing baseline standards for analysing these methods, a set of candidate simulations and the associated quantification of performance, and results of the method applied to the AIS data described in Section 2.8.

3.1 Introduction

As described in Chapter 1, this thesis has used Automatic Identification system (AIS) data augmented with a copy of the IHS Markit World Register of Shipping database to develop tools and techniques to understand maritime behaviour.

AIS is often believed to provide ground truth [31, 112]. These studies are all in the context of single AIS receiver (or group of receivers) in a managed network by the researchers for a contained small area ($40,000\text{km}^2$ less than 0.01% of the area of the ocean).

Looking at the specification of AIS, as described in Section 2.7, a set of assumptions can be formulated that will perform ship tracking using AIS data. These assumptions include:

- Each vessel has a unique Mobile Maritime Service Identity (MMSI).
- Each vessel reports its position and identity at a set rate.
- All AIS transmitted messages are correct and truthful.

- An AIS receiver receives all transmitted messages.
- All AIS received messages are correct.

Using these assumptions and selecting a single vessel, filtered using the Mobile Maritime Service Identity (MMSI), you would get a track similar to that shown in Figure 3.1.

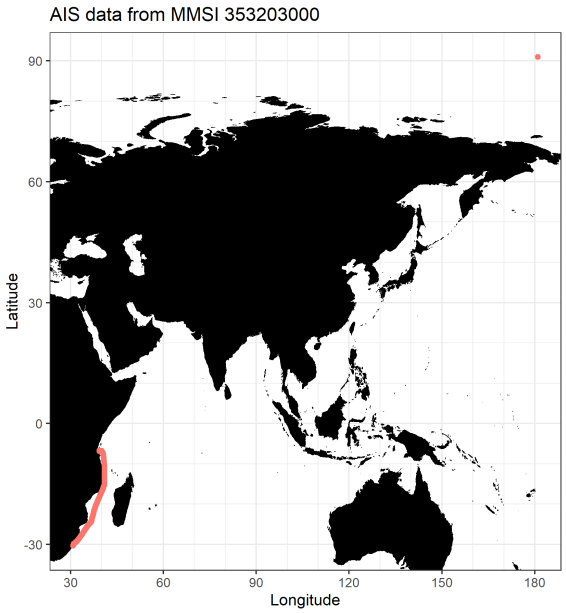


Figure 3.1: MMSI 353203000 with erroneous ($91°N$, $181°E$) messages

However, AIS often includes erroneous reports which means that just joining the measurements will look that shown in Figures 3.2 and 3.9.

Clearly a ship cannot repeatedly travel between the Pacific and Atlantic oceans every hour! We therefore need a way of filtering out the erroneous data points. This is where a tracker comes into play.

From the exploration of the datasets in Chapter 1, it can be seen in the distance vs. time, distance between consecutive observations and time between consecutive observation plots that there are many duplicate "vessels" using the same MMSI. The global plot of observations per MMSI (Figure 2.30) shows that even when a dataset is constrained to provide a single observation per hour for 15 days (a total of 360 hours), there are many examples of more than one vessel reporting on the same MMSI. Figures 3.11, 3.12 and 3.13 show, for each dataset, that this phenomenon is not limited to hourly snapshot

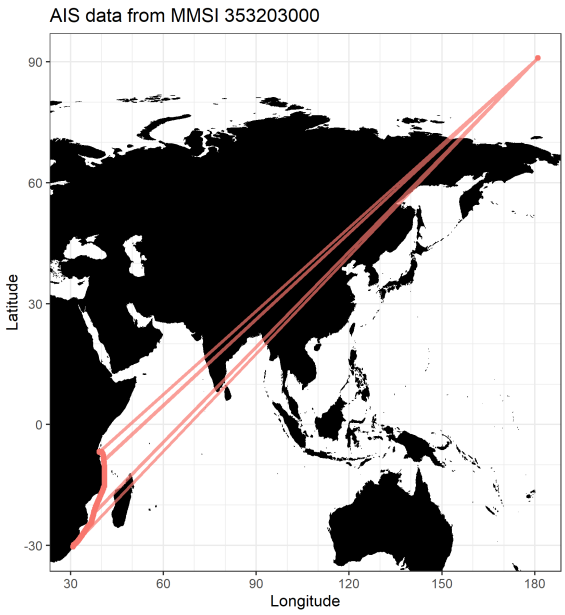


Figure 3.2: MMSI 353203000 with erroneous ($91°N$, $181°E$) messages with observations joined.

data. Consequently, the data in their raw form does not provide useful insight into vessel behaviour for all vessels.

AIS data therefore needs to be "cleaned" to provide users with an unambiguous picture of their region of interest. Such a picture would enable an analyst to clearly distinguish vessels leaving the region of interest (because they did) rather than a duplicate MMSI reporting outside the area of interest.

The examples shown highlight some of the issues encountered with the assumptions when using AIS information. These issues include:

- There are many examples of multiple vessels using the same MMSI, highlighted in Figure 2.32 which leads to a reporting rate of a given MMSI being greater than that of the specifications (e.g., those shown in Figure 2.30).
- Due to network limitations the received data rate drops as vessels travel in open ocean.

This chapter outlines the use of a multiple target tracker and its use to clean this data and split these erroneous multiple tracks from multiple vessels.

3.2 Quantification of performance

The following simulated cases are based on the behaviours evident in the AIS datasets as described in Section 2.8.4. Some behaviours described relate to vessels changing MMSI number, this section modifies these behaviours to specific versions where there are multiple vessels reporting on the given MMSI.

This section will use two metrics to assess performance, Generalised Optimal Sub-Pattern Assignment (GOSPA) [101, 108, 118] and Single Integrated Air Picture (SIAP) [139, 92]. The GOPSA metric is calculated at each time step, returning an overall multiple-tracks to multiple-ground truth missed distance and the distance the track is away from the truth [108] (for a given distance metric, in these simulations, Mahalanobis distance was used). This has two properties; p the exponent, for outlier sensitivity, and c, the cut off distance, for cardinality penalty. The GOSPA metric is a generalisation of the OSPA metric which is divided into four components; distance, localisation missed detection, and false alarm. The distance for the GOSPA metric is a combination of localisation, missed detection, and false alarm. GOSPA produces a distance for each time step a track exists. To summarise multiple simulations, this distance per time step is aggregated to the mean distance over a simulation The goal is to minimise this distance. The SIAP Metric computes attribute measures such as ambiguity, completeness, and spuriousness [10, 35, 90]. The main attribute metric used for assessing the performance of the simulations in this section is the completeness attribute which is the percentage of ground truth covered by a track. The completeness was produced twice, once with all ground truth objects, to assess the overall performance and again where the ground truth was only the true vessel's ground truth.

3.2.1 Normal behaviour of a single vessel reporting on a single MMSI

Figures 3.3, 3.4 and 3.5 show example scenarios of a single vessel and depict the ground truth, measurements and resultant track with various levels of measurement noise (q = 10, 1000, and 10000 respectively) to simulate the noise on the AIS positional messages. Table 3.1 lists the simulation parameters for the cases and the GOSPA distance metric and the SIAP Completeness metric.

The simulations exemplified by Figures 3.3, 3.4 and 3.5 and the associated metrics are displayed in Table 3.1, rows 1-3. These cases were run 100 times and the mean metrics are displayed. The SIAP completeness metric has a completeness of 1 for $q = 10$ and 10000

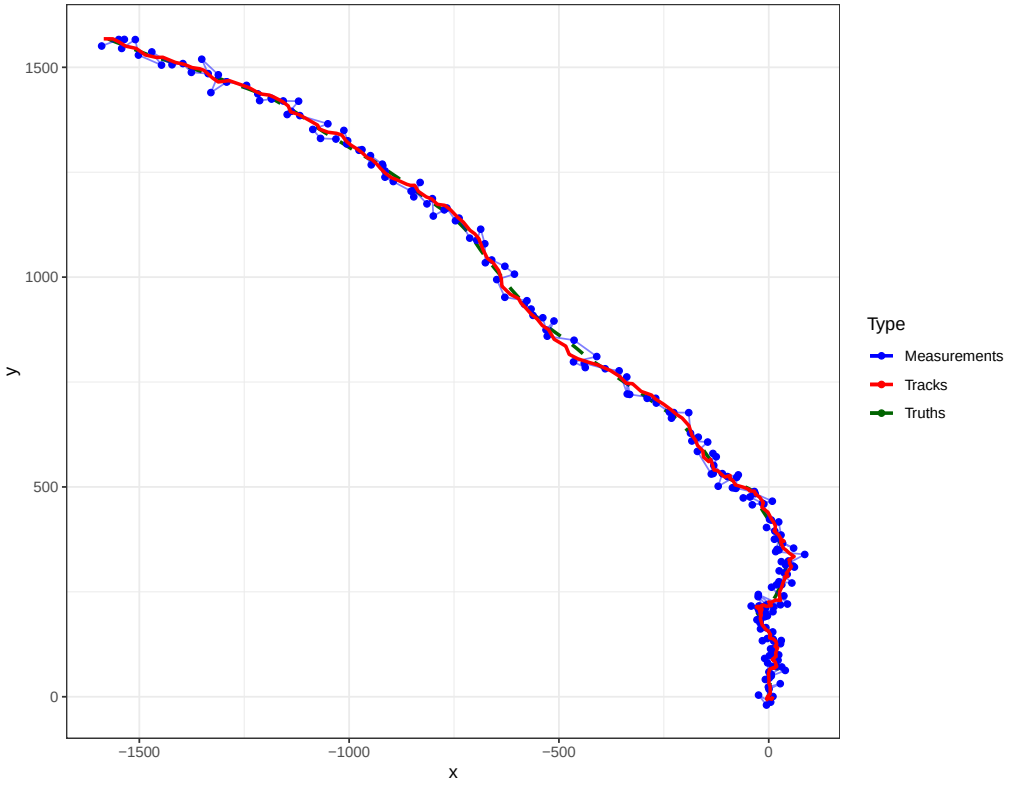


Figure 3.3: Example simulation of a single target with measurement noise $q = 10$.

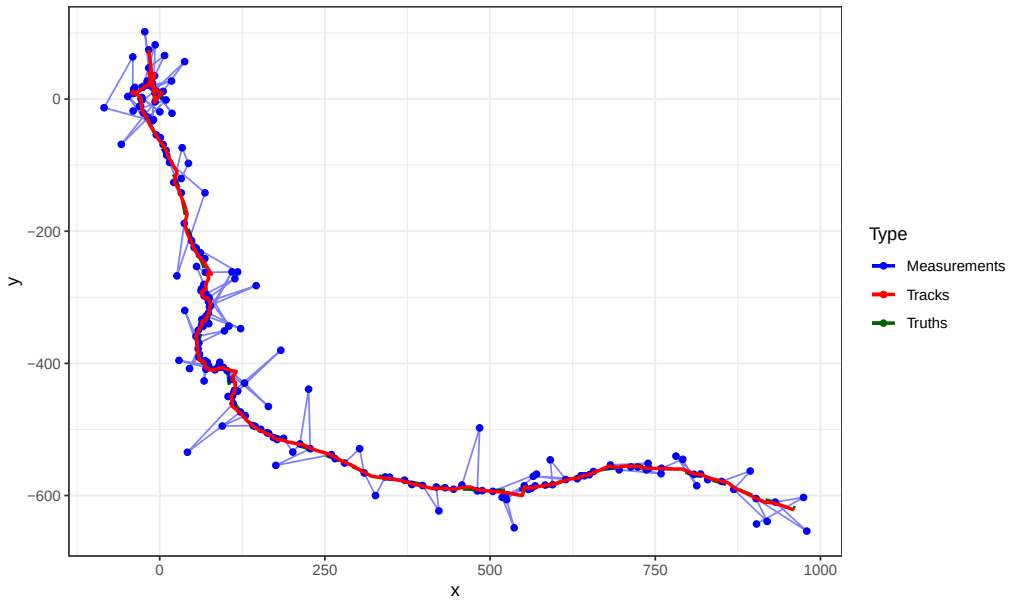


Figure 3.4: Example simulation of a single target with measurement noise $q = 1000$.

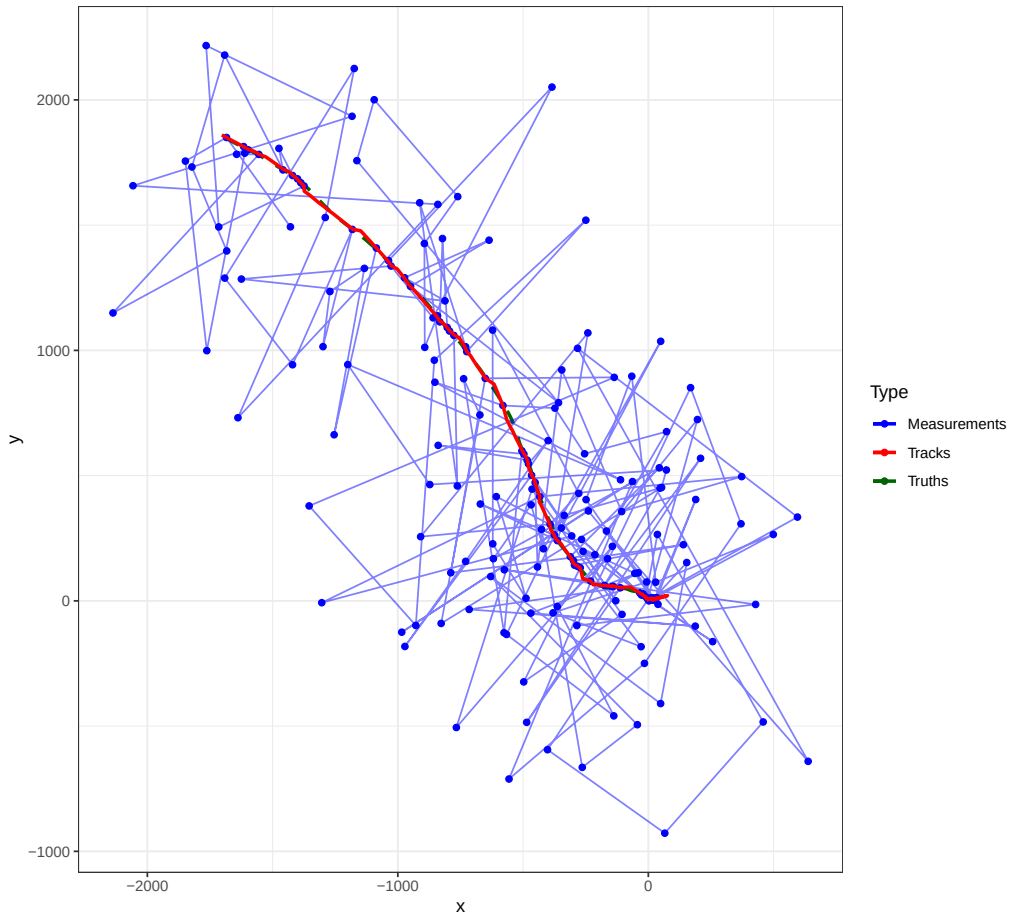


Figure 3.5: Example simulation of a single target with measurement noise $q = 100000$.

and 0.97 when $q = 100000$ which corresponds to 97% of the truth was covered by the track.

The set of simulations described in Table 3.1 increase the number of other objects in the scenario, v additional vessels, b additional stationary objects and p represents the probability of the measurements being generated from the true vessel. Figure 3.6 provides an example of a more complex scenario. The results show that generally all truth objects (true vessel plus v additional vessels and b additional objects) have a high completeness C_A while the completeness of the true vessel C_T is dependent on the amount of measurements that were generated by the true vessel's ground truth p. Figure 3.7 shows the comparison of the total number of measurements generated by the true target against the completeness metric for all truth objects (C_A) and depicts the high completeness of truths being covered by tracks. Figure 3.8 shows the relationship between the total number of measurements generated from the true vessel and the completeness of the tracks to the true vessel. The figure shows a high correlation between number of measurements generated by the true vessel and the amount of the true vessel truth that was covered by the track. Additionally, the figure shows that the full true vessel track can be inferred from as little as 50% of measurements (where the measurements are generated by the true vessel).

These simulations show that the tracked vessels can be tracked while more than one vessel is reporting on the MMSI of the true vessel.

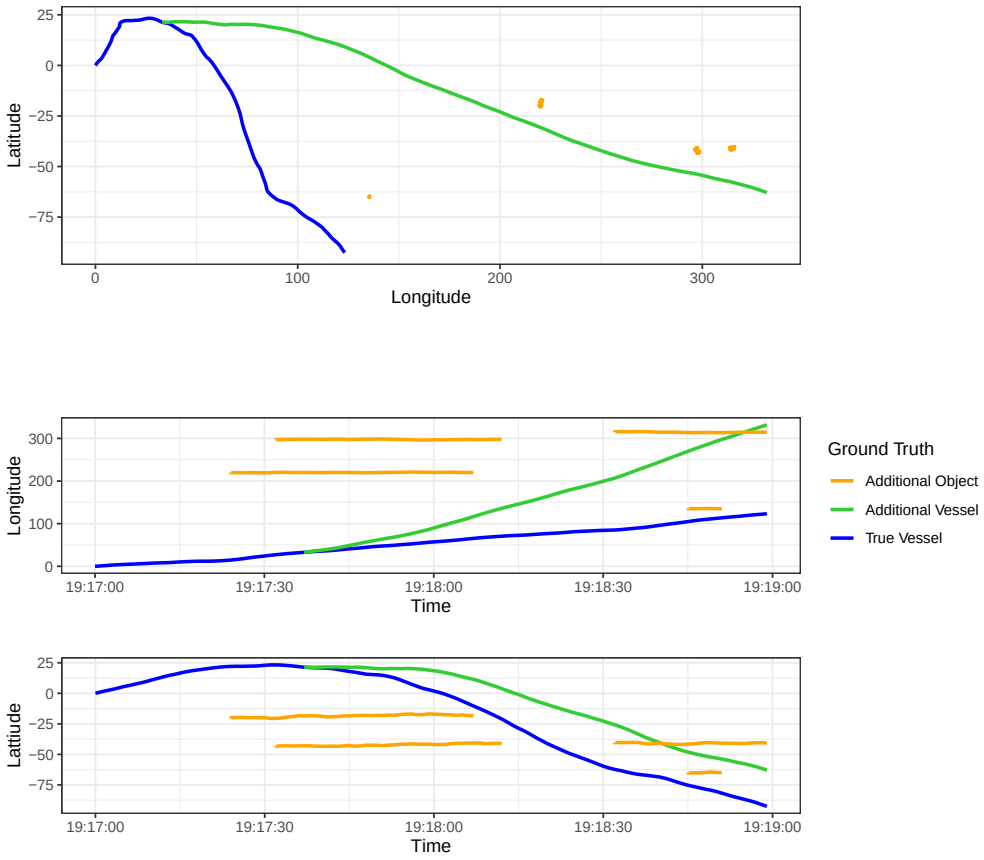


Figure 3.6: Example simulation of a True vessel (blue), 1 additional vessel (green), in this example, spawning from a point along the true vessel trajectory, and 4 additional objects (orange).

n	v	b	S	p	q	C_A		C_T		d	
200	0	0	100	1	10	1	(0)	1	(0)	8.6565	(2.4703)
200	0	0	100	1	1000	1	(0)	1	(0)	9.8469	(0.5682)
200	0	0	100	1	100000	0.97	(0.02)	0.97	(0.02)	94.9626	(9.7999)
200	1	0	100	0.75	100	0.96	(0.02)	0.99	(0.02)	86.1857	(22.7934)
200	1	0	100	0.5	100	0.96	(0.03)	0.95	(0.03)	91.8896	(20.1351)
200	1	0	100	0.25	100	0.95	(0.02)	0.93	(0.04)	90.3758	(26.7675)
500	2	0	100	0.75	100	0.95	(0.05)	0.97	(0.02)	88.1053	(20.1246)
500	2	0	100	0.5	100	0.84	(0.08)	0.9	(0.02)	52.1368	(39.6485)
500	2	0	100	0.25	100	0.92	(0.1)	0.83	(0.03)	75.998	(33.8748)
500	2	10	100	0.75	100	0.86	(0)	0.95	(0.09)	68.4105	(33.1738)
500	2	10	100	0.5	100	0.89	(0.07)	0.86	(0.09)	46.7654	(24.3875)
500	2	10	100	0.25	100	0.87	(0.04)	0.77	(0.04)	27.1339	(18.4932)
500	10	10	100	0.95	100	0.89	(0.08)	0.99	(0.07)	29.4788	(31.0242)
500	10	10	100	0.75	100	0.91	(0.07)	0.92	(0.06)	30.2407	(17.5357)
500	10	10	100	0.5	100	0.88	(0.08)	0.86	(0.07)	19.6023	(18.6548)
500	10	10	100	0.25	100	0.86	(0.08)	0.4	(0.01)	40.5771	(23.5744)
500	10	10	100	0.25	100	0.78	(0.1)	0.18	(0.08)	23.569	(43.2493)

Table 3.1: Results of 20 scenarios each simulated over S runs of n time steps with 1 true vessel, v additional vessels, maximum of b additional objects, and p probability of measurements generated by the true vessel. The results SIAP completeness for all truths C_A, completeness for the true vessel C_T, GOSPA distance d.

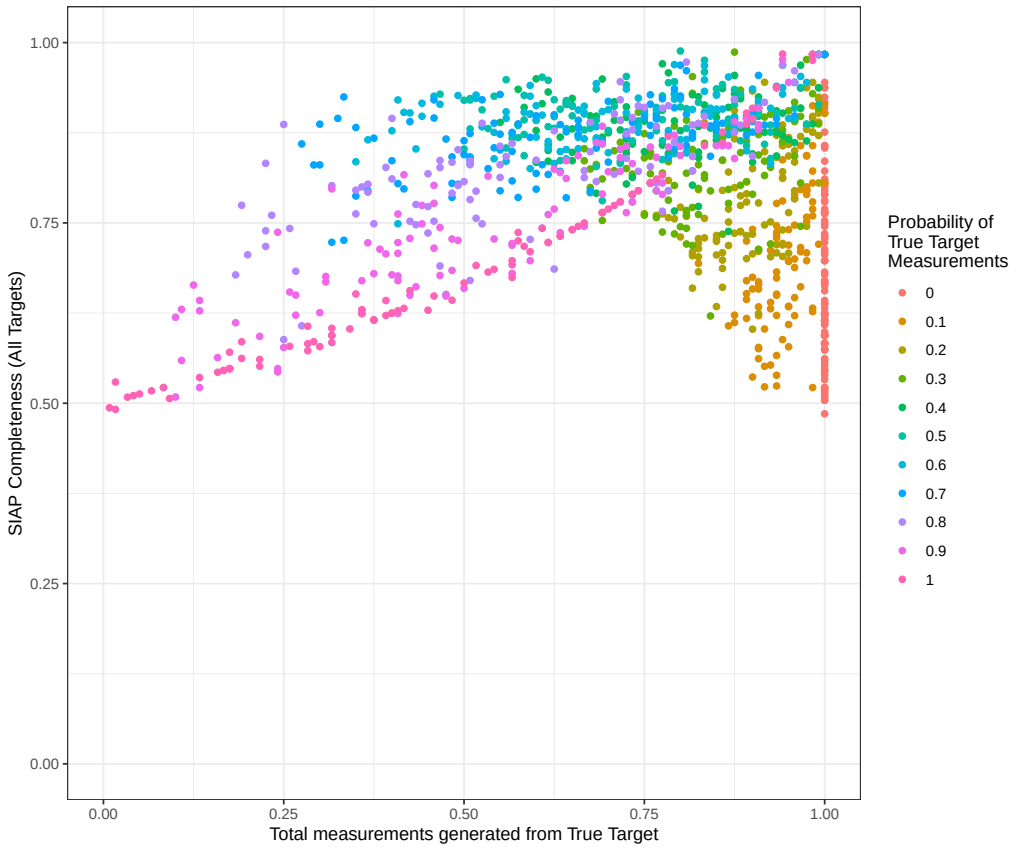


Figure 3.7: Simulation results of SIAP Completeness Metric (on all tracks and all vessels) vs the percent of actual measurements generated from true vessel for n probability of detecting true vessel over other.

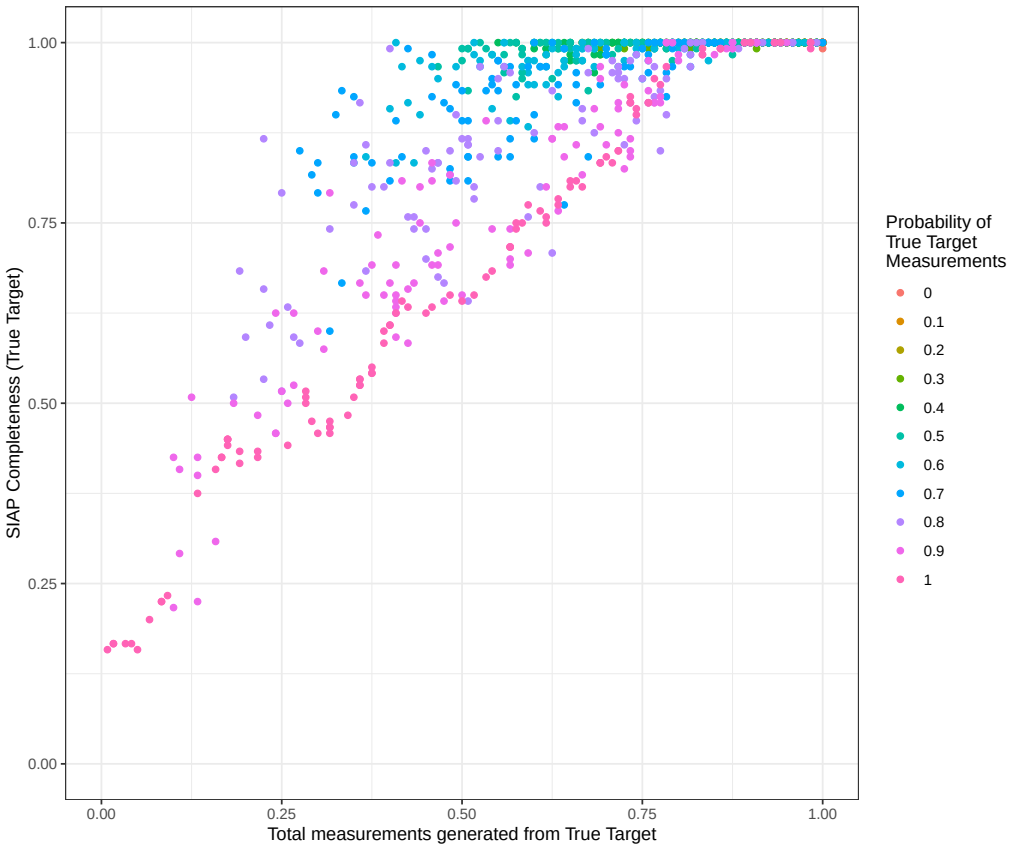


Figure 3.8: Simulation results of SIAP Completeness Metric (on all tracks and only the true vessel) vs the percent of actual measurements generated from true vessel for n probability of detecting true vessel over other.

3.3 Applying the Tracker to AIS data

To begin each object was considered to be linearly Gaussian (the latitude and longitude were converted to UTM coordinates). The position of the object was measured at every Δt hours. These measurements are deemed to be considered 'accurate' since they are derived from GPS (since the error associated with GPS measurements is very small) and likely to have already been filtered (by the AIS provider). Any imprecise measurements are assumed to be erroneous or malicious.

The parameters of the tracker applied to the datasets based on the results of the quantification of performance based on the results shown in Section 3.2 are

- Transition model: Ornstein-Uhlenbeck process $1m$ process noise with decay coefficient, $K = 2 \times 10^{-3}$.
- Measurement model: Linear model with $10m$ noise.
- Prior at $[0, 0, 0, 0]$ with noise in the position of $10m$ and $1ms^{-1}$.
- Global Nearest Neighbour data association was used with a Mahalanobis distance of 10 standard deviations as the gating region.
- Tracks are initiated on Measurements.
- Tracks are deleted from the active tracks when they have not been seen for 12 hours.

The latitude and longitude are converted to UTM coordinates and are set as x and y. Using a constant velocity model, the state becomes

$$\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_k + \boldsymbol{w}_k \tag{3.1}$$

where $\boldsymbol{x}_k = \begin{bmatrix} \lambda \\ \dot{\lambda} \\ \varphi \\ \dot{\varphi} \end{bmatrix}$, $\boldsymbol{F} = \begin{bmatrix} 1 & \delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q})$, and $\boldsymbol{Q} = \begin{bmatrix} \frac{1}{4}\delta t^4 & \frac{1}{2}\delta t^3 \\ \frac{1}{2}\delta t^3 & \delta t^2 \end{bmatrix} \sigma_a^2$.

At each time step, a true measurement of the object is made (if noisy). Let the noise of the measurement $\boldsymbol{v}_k$ be normally distributed σ_z.

$$\boldsymbol{z}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}_k \tag{3.2}$$

where $\boldsymbol{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R})$, and $\boldsymbol{R} = \begin{bmatrix} \sigma_z^2 \end{bmatrix}$.

For each object, the initial starting state was initialised as the first observed state

$$\hat{\boldsymbol{x}}_{0|0} = \begin{bmatrix} \lambda \\ 0 \\ \varphi \\ 0 \end{bmatrix} \tag{3.3}$$

The position is assumed to be exact, a zero covariance was assigned.

$$\boldsymbol{P}_{0|0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.4}$$

Alternatively, to include some noise in the initial state, the covariance could be

$$\boldsymbol{P}_{0|0} = \begin{bmatrix} \sigma_\lambda^2 & 0 & 0 & 0 \\ 0 & \sigma_{\dot{\lambda}}^2 & 0 & 0 \\ 0 & 0 & \sigma_\varphi^2 & 0 \\ 0 & 0 & 0 & \sigma_{\dot{\varphi}}^2 \end{bmatrix} \tag{3.5}$$

This demonstrates that when using just the Kalman filter to provide an inlier/outlier classification, it is important to start on an inlier (i.e., the true track) rather than an outlier (i.e., an error message (91,181)).

Track management provides a way to track both the inliers and outliers. This was done by initiating a new Kalman filter on any data point that does not associate with any existing Kalman filters. This can be seen in Figure 3.10.

To formulate this bank of Kalman filters, a track management system was needed that can initiate a new Kalman filter, manage existing Kalman filters, and delete tracks that

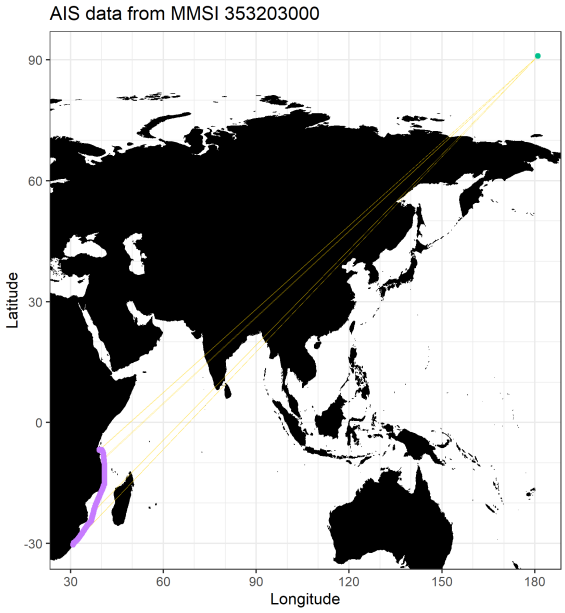


Figure 3.9: MMSI 353203000 with erroneous ($91°N$, $181°E$) messages with observations tracked by a Kalman filter. The yellow lines denote the raw data and the purple line represents the tracked vessel and the blue points represent the discarded observations.

have not received data for a particular length of time. The following section will introduce the track management system that can cope with this bank of Kalman filters.

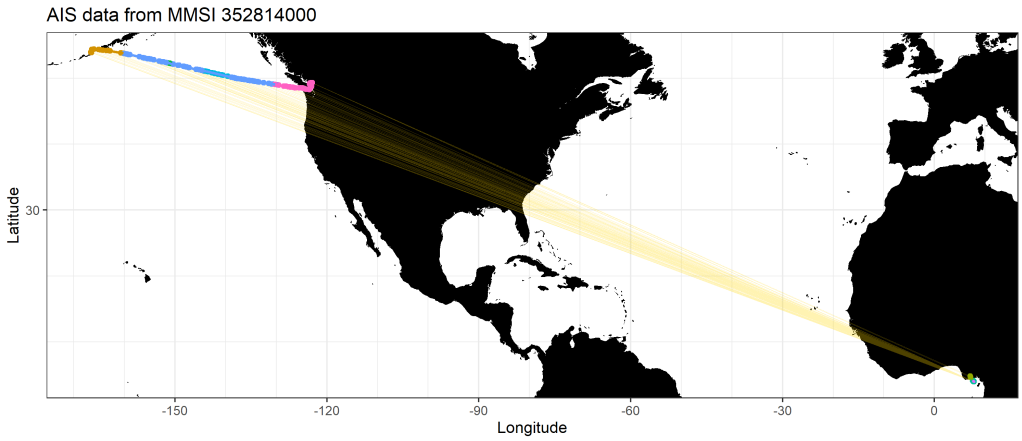


Figure 3.10: Result of initiating a new Kalman filter on AIS positions further away than existing Kalman filters.

3.4 Results

Figures 3.11, 3.12 and 3.13 show a set of MMSI numbers being used by more than one vessel for each dataset. Figure 3.16 shows the results of the multiple target tracker applied to the global dataset from Figure 3.13. There is a significant improvement over the original data concluding that disambiguating MMSI numbers at the global scale is useful.

Figure 3.15 shows the results of the multiple target tracker applied to the North Atlantic dataset from Figure 3.12. Like the global dataset results, the improvement of the North Atlantic dataset results over the initial data is evident. This shows that not only does the multiple target tracker work with the fixed interval (hourly snapshot) data of the global dataset but can accommodate variable time interval (2 – 10 seconds) data in the North Atlantic dataset.

Figure 3.14 shows the results of the multiple target tracker applied to the Merseyside data from Figure 3.11. There is an improvement over the original data but in appearance, it is not as successful as that of the visual appearance of the global dataset or the North Atlantic dataset. This is a result of the smaller geographical area of the Mersey Estuary

and hence the tracks have not been deleted from the tracker as the distances are still within the error covariance threshold.

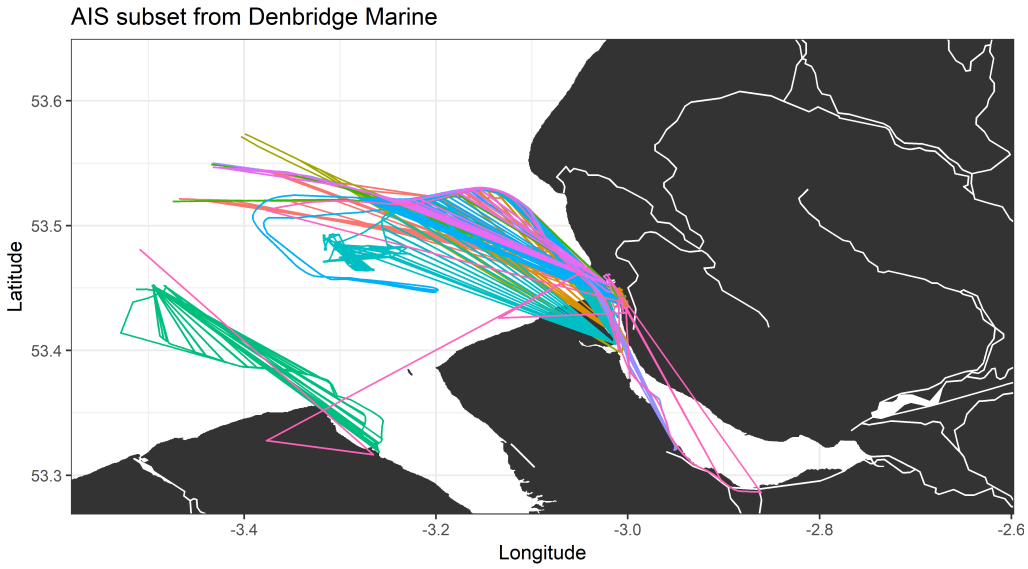


Figure 3.11: Subset of data from a single AIS receiver where each MMSI point has been joined into a path.

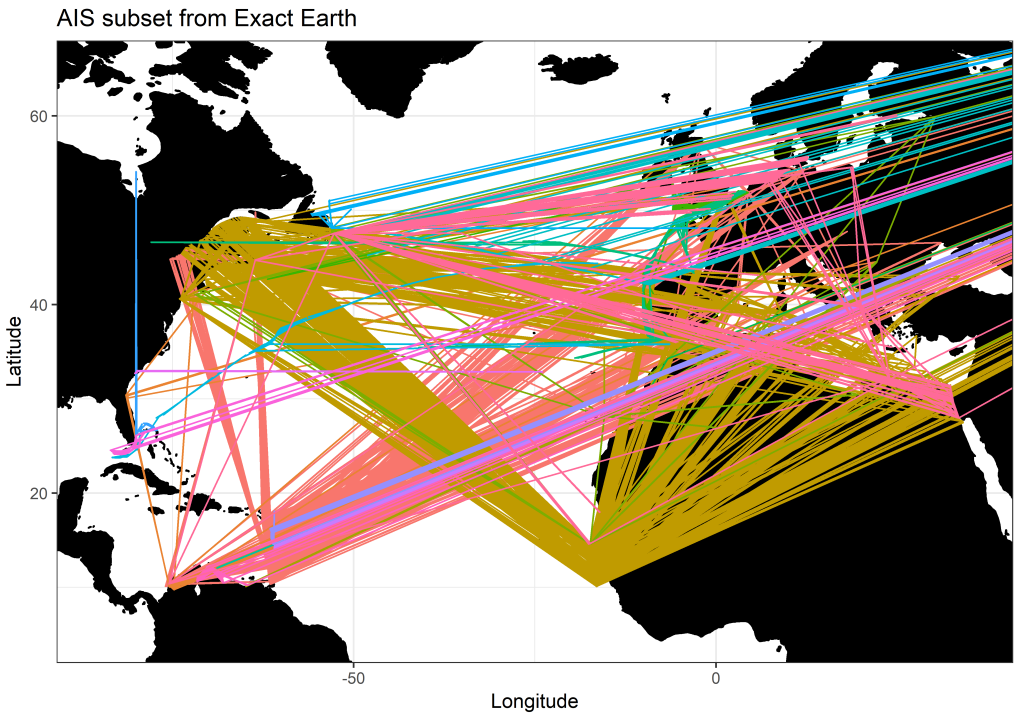


Figure 3.12: Subset of interesting MMSIs from the North Atlantic dataset.

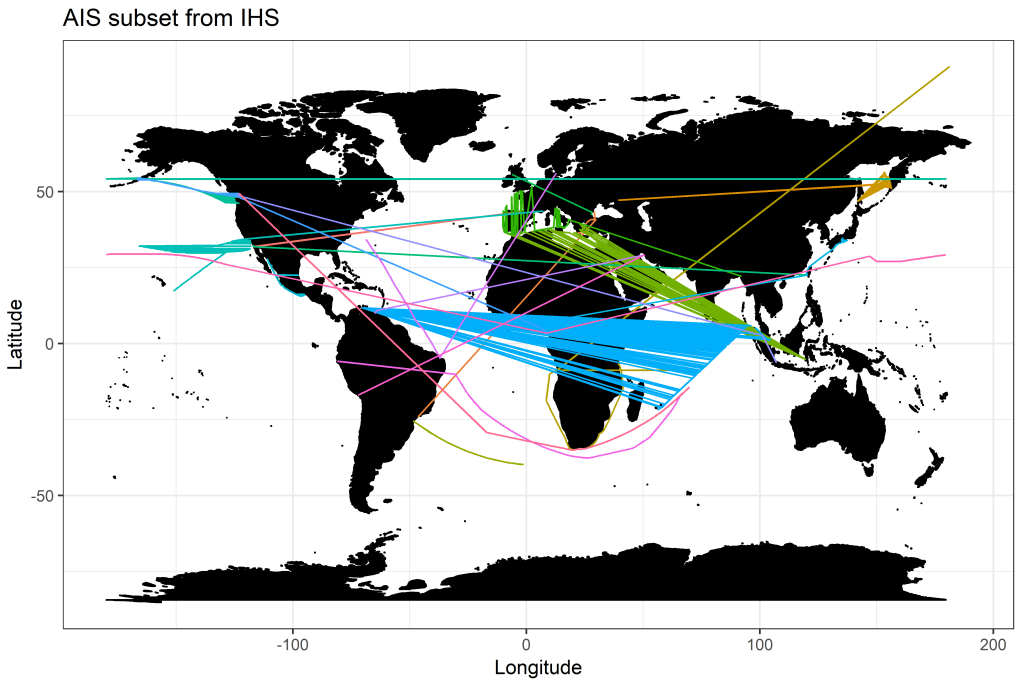


Figure 3.13: Subset of interesting MMSIs from the Global dataset.

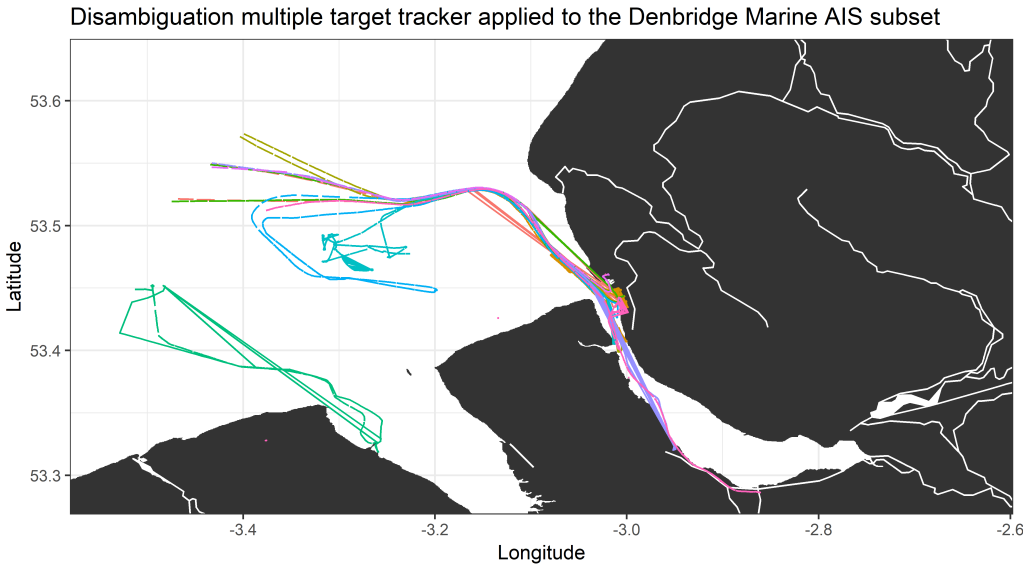


Figure 3.14: The result of applying the disambiguation process to the Fort Perch Rock dataset.

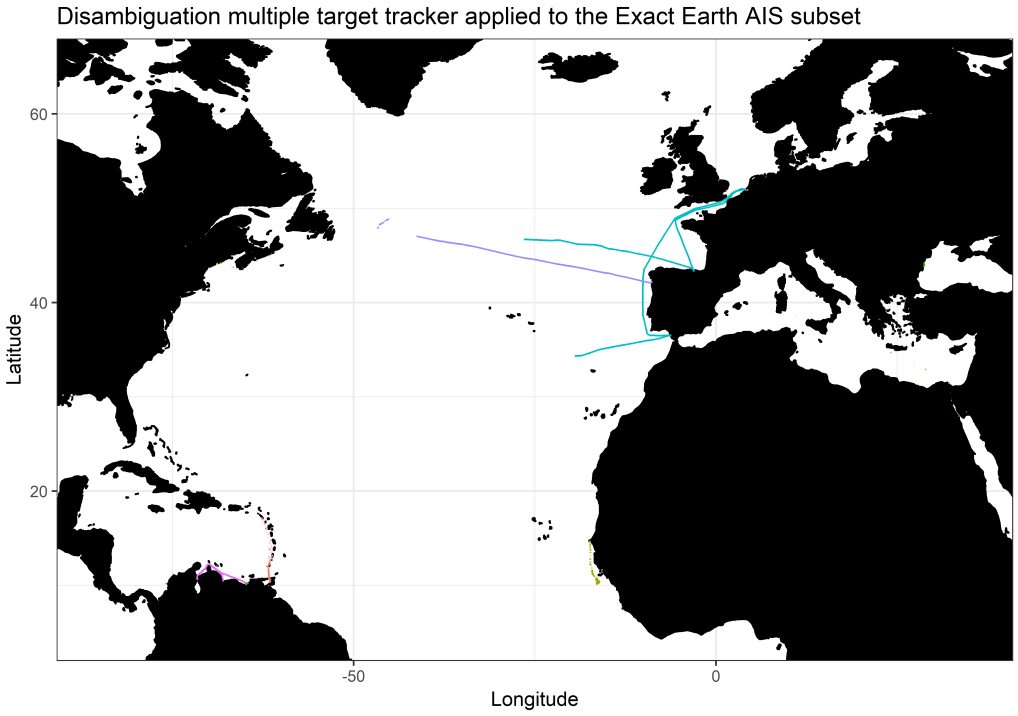


Figure 3.15: The result of applying the disambiguation process to the North Atlantic dataset.

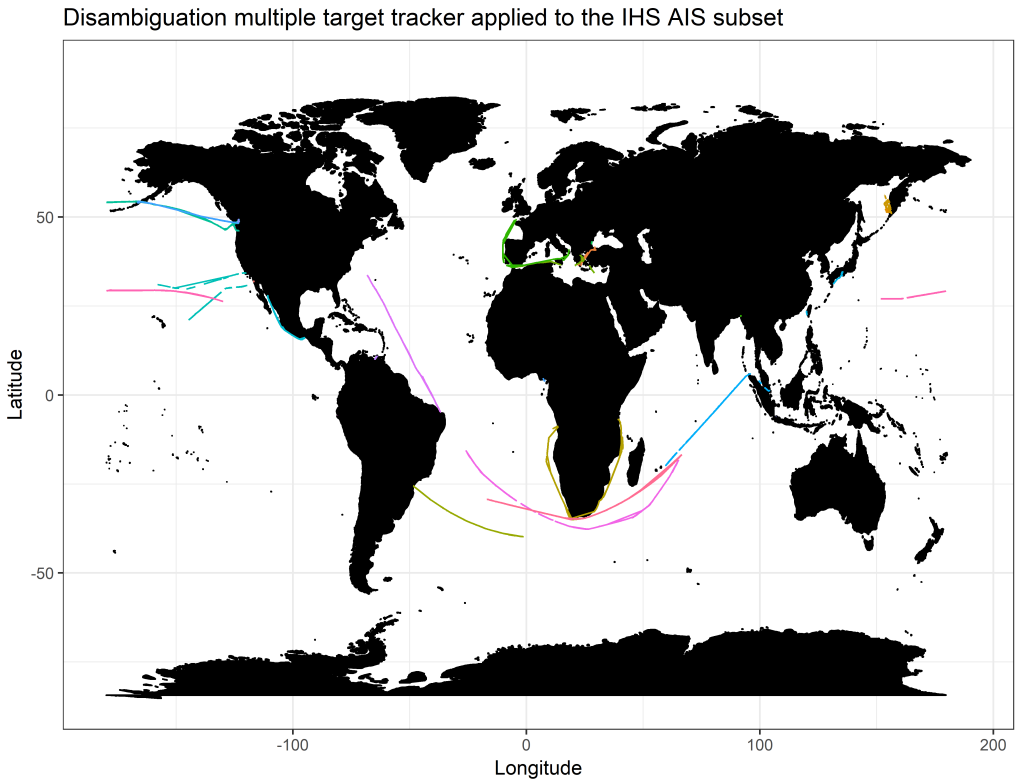


Figure 3.16: The result of applying the disambiguation process to the Global dataset.

3.5 Discussion

The method proposed in this chapter provides a more intuitive view of the data that can be presented to a human operator.

There are some areas in which the tracker could be improved. Future improvements could include:

- The motion model could be improved. The results in Section 3.4 used an Ornstein-Uhlenbeck mean reverting stochastic process transition model in only latitude and longitude. The AIS messages also provide course over ground (COG), and speed over ground (SOG), which can be used to provide more insight into the measurement state while adding complexity by increasing the dimensionality of the state.

- Along the same lines as improving the motion model, multiple models can be introduced such that they can describe the different possible dynamic states a vessel can inhabit. An Interacting Multiple Model, IMM [94], would be able to model when a vessel is moving (using a constant velocity or Ornstein-Uhlenbeck process) and when a vessel is stationary (using a random walk model) by switching between the models.

Chapter 4

Single Ship Analysis

This chapter focuses on developing methods to analyse vessel tracks on a ship by ship level. The research within the chapter develops methods based on the disambiguated data from Chapter 3 and focusses on tracklet joining within a MMSI and reflagging (the joining of tracklets between different MMSIs), and ship stopping.

With the output from Chapter 3, additional information can be generated from the tracklet data that could not be calculated when more than one vessel was sharing a MMSI number.

This chapter discusses methods related to individual ships in the larger maritime picture. Techniques that can extract further details from the data are introduced that are able to provide a human operator with a greater wealth of information about a particular vessel.

During this study a number of techniques are employed, including:

1. *Automatic tracklet joining*: This section looks at the joining of tracklets from the same MMSI and specifically analysing the probability a vessel changed from reporting on one MMSI to another MMSI.

2. *Analysis of stopped vessels*: This section looks at detecting when a vessel has stopped to provide a list of geographical positions vessel's stop.

4.1 Track Joining and Reflagging

Tracklet joining discussed in Section 2.2 the literature, (e.g., [136]) focussed on only predicting forward from the first tracklet to the start of the second tracklet to generate

the cost of joining a tracklet pair. Figure 4.1 shows the ROC curve comparing the difference between using the forward prediction and using both the forward and backward predictions. The figure indicates that the combination of the forward and backward predictions outperforms the forward prediction.

This is extended to a set of simulations at different tracklet pair counts and the associated ROC curve can be seen in Figure 4.2. The simulation generated a set of broken tracks for each of a set of ground truths. The number of possible tracklet pairs was the complete set of all tracklet to tracklet joins in the simulation (ignoring temporal and spatial gating). The figure shows for a low number of tracklet pairs, the classification accuracy is extremely high (and in the case of the trivial 1 tracklet pair, the accuracy is 1 where the ROC curve passes through (0,1)). As the number of tracklet pairs increases, the performance decreases which was due to the number of different possible assignments in each simulation.

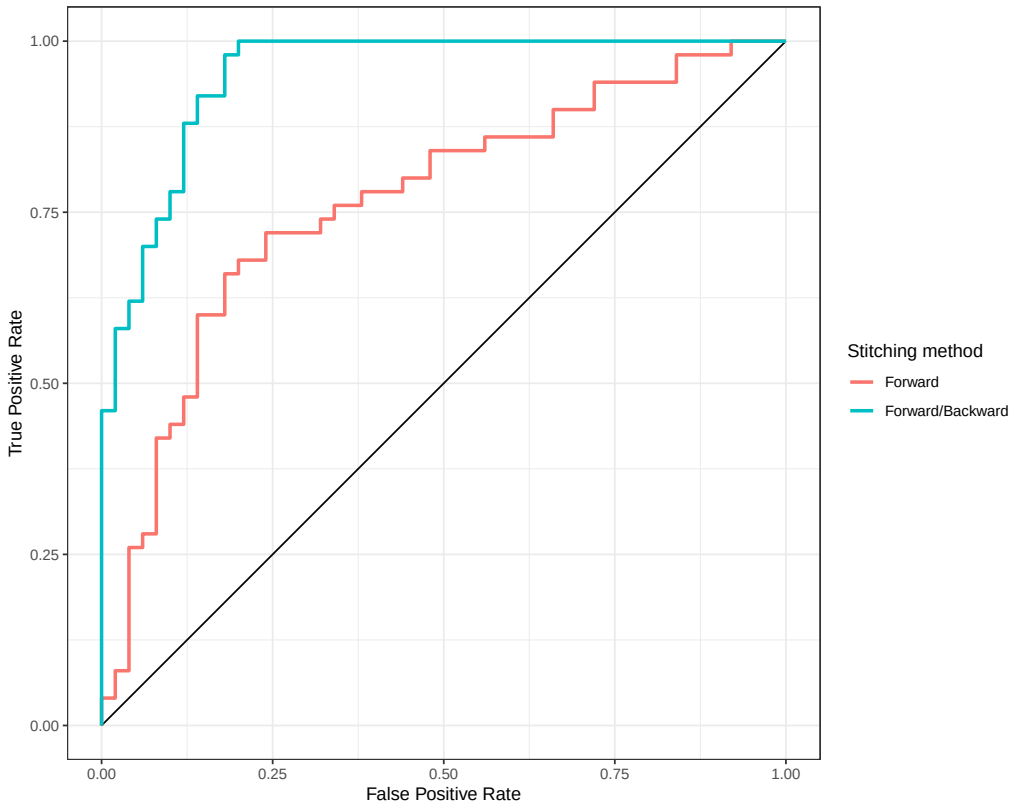


Figure 4.1: The ROC curve depicts the difference between using the predicting forward only, and the combination of using both forward and backward predictions.

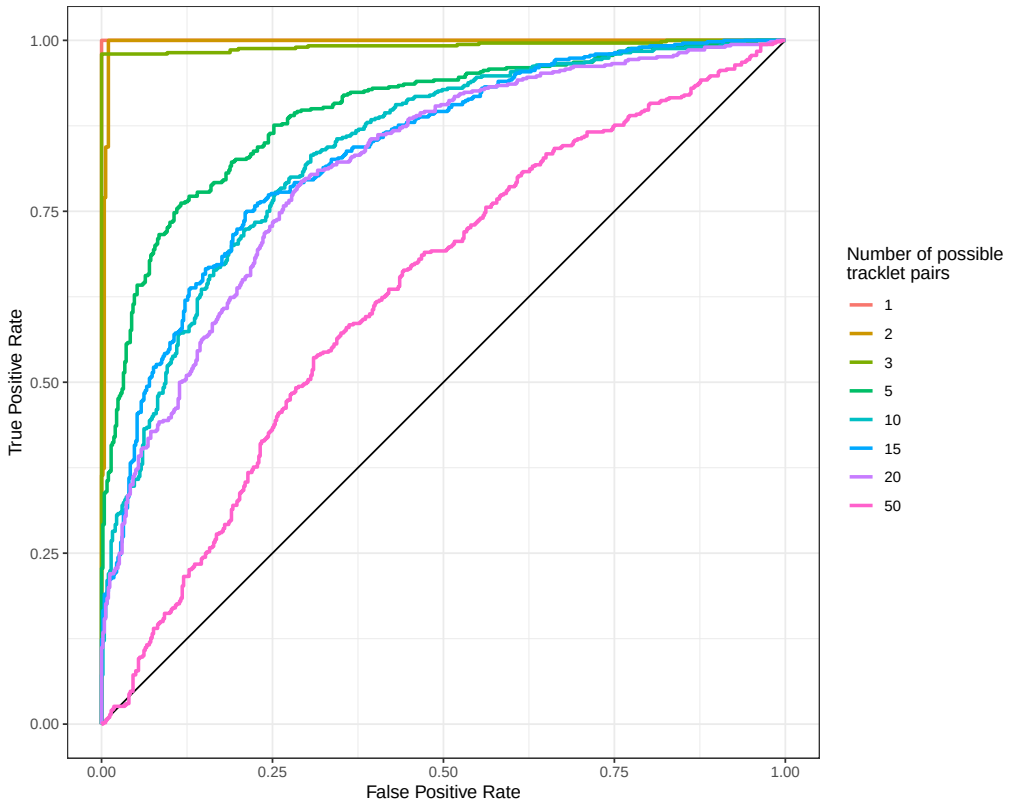


Figure 4.2: The ROC curve depicts the difference between the number of possible tracklet pairs and their assignment using the combination of forward and backward predictions to generate the joining cost.

4.1.1 Results for Reflagging

The aim of reflagging is to collate a list of likely suspects of vessels changing MMSI. This list is passed to a human operator for further enquiry. As mentioned in Chapter 1, there are a number of commercial entities that keep track of all vessels changing MMSI (as legal changes are allowed). These databases can take a few weeks to be generated and circulated so, the outputted list here gives operators more instant notice that a vessel appears to be reflagging. These databases of reflagging can be visualised as shown in Figure 2.5c and allows the human operator to allocate new identifiers to the MMSIs in the reflagging event (Figure 2.5d).

Tracklet 1 ID	Tracklet 2 ID	Additional Info...
100000001-1	100000004-3	More than 1 option see row 5
100000002-1	100000001-3	
100000001-2	100000005-1	
100000003-1	100000002-1	
100000005-3	100000004-3	More than 1 option see row 1

Table 4.1: A list of reflagging events presented to a human operator

4.2 Ship Stopping

4.2.1 Utilising a by-product of the multiple target tracker

To calculate if a vessel is stationary, the vessel's speed is needed. The speed of a vessel was calculated in the disambiguation process for each observation. Therefore, the velocity components of the state can be used to estimates of the Kalman filter from the multiple target tracker for a given track. The vessel speed was calculated from the velocity components and a threshold of 5 knots was applied making the assumption that a vessel travelling below this threshold is near stationary or has stopped moving.

Each stationary occurrence is collated with a set of summary statistics. These will include the location (The mean location from all the stationary locations), and the time extent of the occurrence.

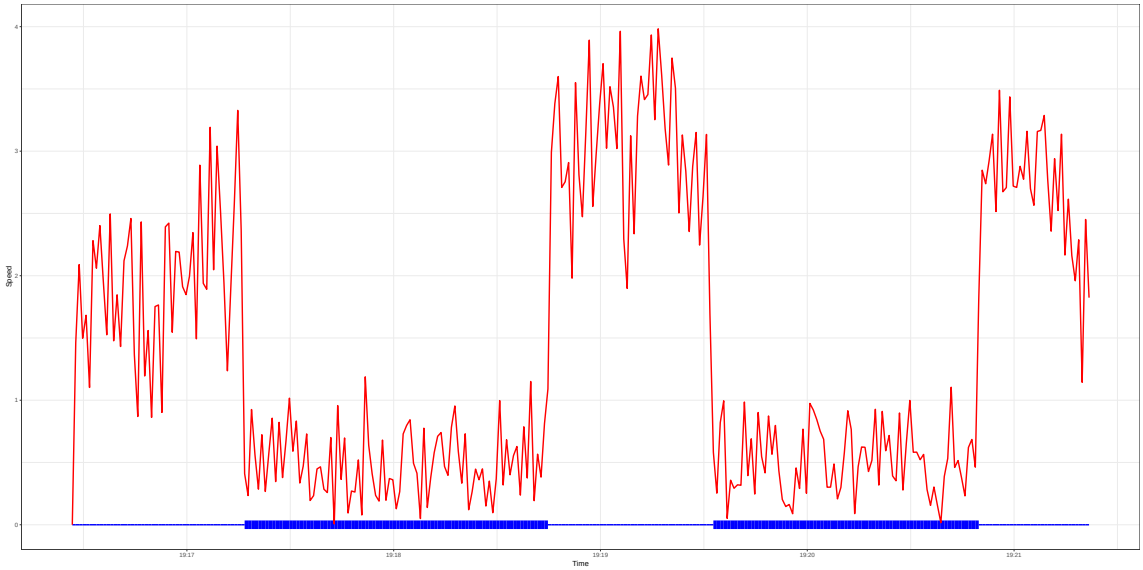


Figure 4.3: Simulation of a vessel moving (thin blue line) and stopping (thick blue line) and the associated speed (red) calculated from the velocity components of the track states.

Figure 4.3 depicts an example simulation where the speed of the vessel is calculated from the velocity components of the track state and forms the scenario for the simulations depicted in Figure 4.4. Figure 4.4 shows that for a simulation where the noise is either generated from a low noise measurement model or a high noise measurement model with the probability of noise shown. This shows that a vessel can be detected to have stopped

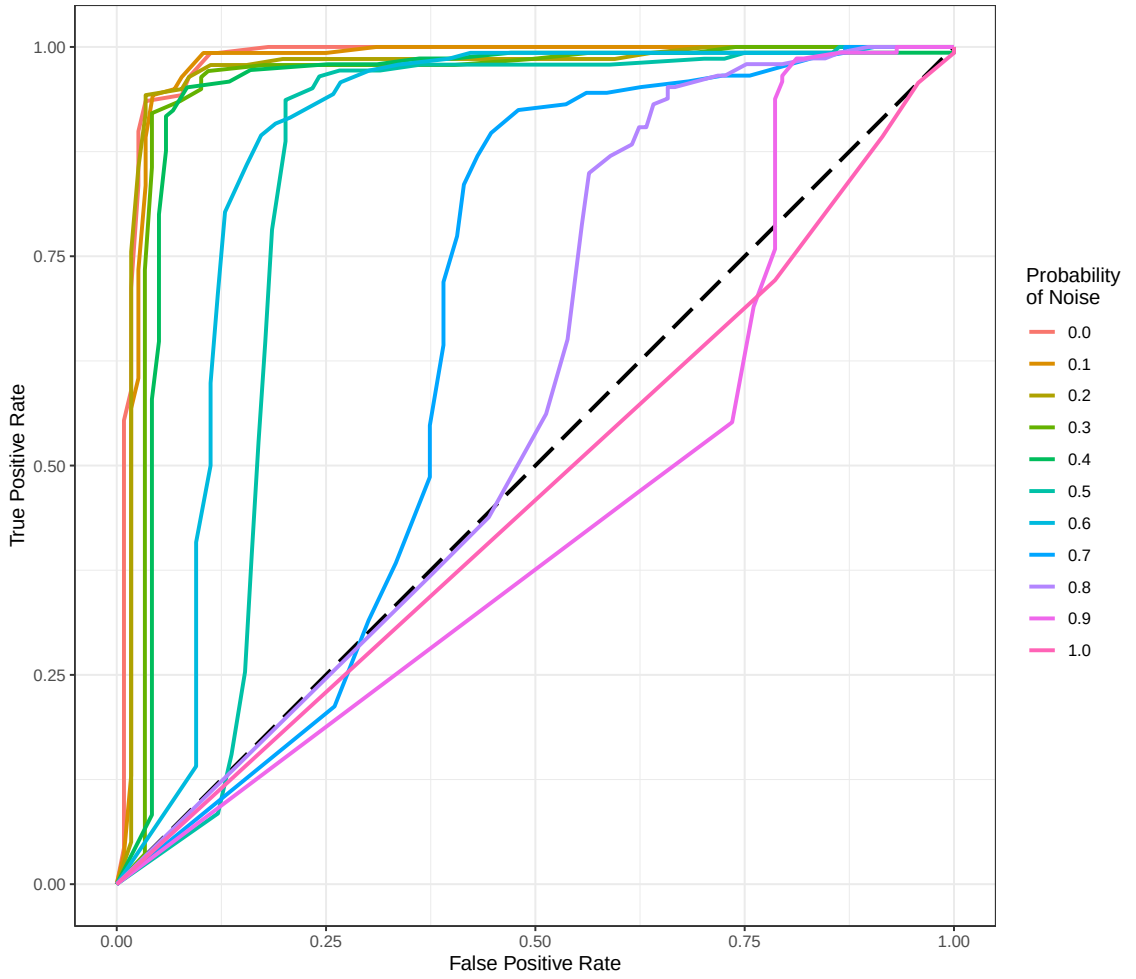


Figure 4.4: ROC Curve depicting the TPR against the FPR for 100 simulations of a vessel moving and stopping (e.g., Figure 4.3) at varying degrees of noisy measurements, where 0 probability of noise refers to the case where there are 0 measurements from a measurement model with high noise ($q = 10000$) and 1 being where the 100% of measurements are from a measurement model with high noise.

where the amount of noisy measurements can be as high as 40%.

Table 4.2 shows the resultant file composed of each vessel's stopped location.

MMSI	Latitude	Longitude	Start	End
123456789	53.583787	-3.700159	2018-08-23 00:23:15	2018-08-23 08:02:24
475209183	2.431654	49.627415	2018-09-02 03:42:54	2018-09-02 04:05:57
275496723	43.89212	35.683832	2018-08-15 23:56:27	2018-08-16 01:03:35
314756812	43.766531	35.901454	2018-08-16 00:45:31	2018-08-16 00:59:47

Table 4.2: An example set of stationary positions

4.2.2 Ports arrivals and departures

The results shown in Table 4.2 can be extended with use of the UN/LOCODE port database (see Section 2.6.3.2), as shown in Table 4.3.

In addition to generating a database of port entries that provide the information of vessel X arrived at port A at time t and left port at time T. For those vessels not in port, each stationary location is allocated with the name and distance to the nearest port and extract its corresponding country to give information that vessel X stopped off the coastline of country Y and time t.

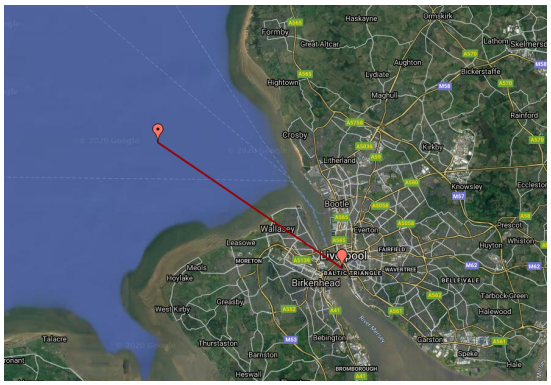
From this detailed set of arrivals and departures, the AIS data has the ability to be broken down into finer detail[1] journeys.

Using this data, events describing the voyage of a ship travelling around the world can be extracted. The voyage is split into a number of journeys. The time between port departure and port arrival can be described as a journey. Every time a vessel stops the database updates the voyage history with a new stopped location. By doing this, the

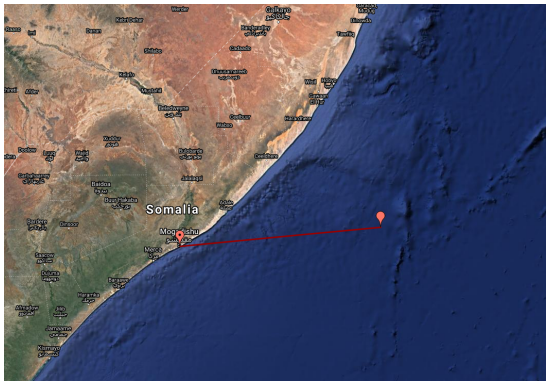
[1]compared to just using the AIS message type 5 destination field.

MMSI	Latitude	Longitude	Start	End	Distance	Port/Country
123456789	53.483787	-3.200159	2018-08-23 00:23:15	2018-08-23 08:02:24	16 km	Liverpool, UK
475209183	2.431654	49.627415	2018-09-02 03:42:54	2018-09-02 04:05:57	481 km	Mogadishu, Somalia
275496723	43.89212	35.683832	2018-08-15 23:56:27	2018-08-16 01:03:35	138 km	Yalta, Crimea
314756812	43.766531	35.901454	2018-08-16 00:45:31	2018-08-16 00:59:47	160 km	Yalta, Crimea

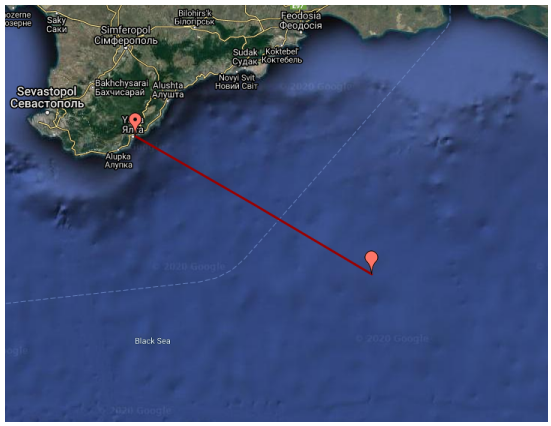
Table 4.3: An example set of stationary positions with details of their nearest ports. This can be visualised to show stopped location and nearest port as shown in Figure 4.5.



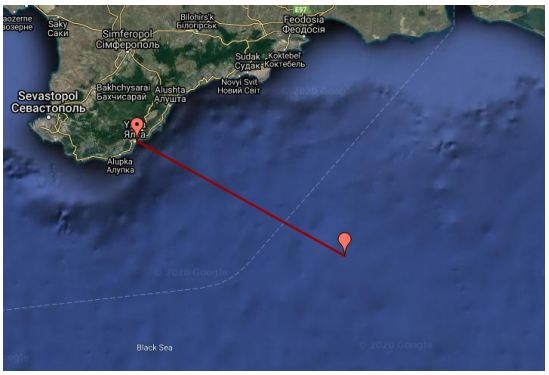
(a) The nearest port (16km) to vessel MMSI 123456789 is Liverpool, UK.



(b) The nearest port (481km) to vessel MMSI 475209183 is Mogadishu, Somalia.



(c) The nearest port (138km) to vessel MMSI 275496723 is Yalta, Crimea.



(d) The nearest port (160km) to vessel MMSI 314756812 is Yalta, Crimea.

Figure 4.5: Visualising the nearest ports to the stopped locations in Table 4.3.

location of the ship is known, which ports it has been to, and where it has stopped outside of a port. This information can help inform human operators on whether the vessel is a security risk.

4.2.3 Discussion

Using this technique, it is possible to see where ships stop at sea. There are many reasons a ship will stop whilst at sea, example include:

- Mandatory drills. A vessel is required by legislation to do a number of safety drills that requires a stop at sea (e.g., man overboard drills and lifeboat drills). This database can help inform whether ships are compiling with their legislative agreement.
- Transshipments at sea. By searching this database over time and location, it is possible to see when two ships are in port or at sea together.
- Unusual activities. The database can be similarly filtered to log vessels that stop a certain distance from the nearest port. This would allow detection of vessels that habitually stop 12 miles off the coast, which is of interest because it could be indicative of illicit activities.

The next stage will be to investigate the use of an interactive multiple model (IMM) tracker. With the implementation of an IMM, within the multiple target tracker framework, in place of a Kalman filter with a single transition model, a set of Kalman filters each with a different transition model can be used. For example, two transition models; one for a vessel that is moving (e.g., constant velocity model) and one for a stationary vessel that is not moving (e.g., constant position model). The velocity thresholding will be simplified by using the mixture probabilities of the mode state from the IMM. The IMM mode index will update when the switching model is chosen as either constant velocity or stationary. This would provide the ability to filter on stationary index to extract the stopped locations.

Chapter 5

Multi-Ship Analysis

This chapter focuses on methods related to improving global understanding of the maritime picture. It describes techniques that can learn behaviours related to a set of vessels to generate alerts to bulk changes of behaviour such as vessels behaving inconsistently and sudden changes in behaviour in a region.

During this study a number of techniques where investigated including:

1. *Behaviour Detection*: Applying LDA and MoU models to the symbolic tracks to infer journey behaviours and vessel types.

2. *Change Point Detection*: This section uses the observations from a given grid cell and uses a change point algorithm on the resultant time series to detect changes in behaviour of vessel frequency over the period in the datasets.

5.1 Behavioural Detection

5.1.1 Latent Dirichlet Allocation analysis of AIS data

Latent Dirichlet allocation is a text analytics algorithm that is used for topic modelling in large corpora of documents. Documents contain a collection of words or phrases that can then derive similar topics over multiple documents. The documents are assumed to be the set of symbolic representations of tracks where the words are the regions generated by the adaptive grid, and the topics are the behaviours of similar vessels.

The collection of symbolic tracks was processed by the LDA algorithm into 10 behavioural clusters. Figures 5.1 and 5.2 show two of these behavioural clusters. The figures show the

likely positions for vessels belonging to the particular behavioural cluster. Behavioural Cluster 2 depicted in figure 5.1 shows vessels predominately in the Irish Sea as well as other smaller spot locations. Behavioural Cluster 8 shown in figure 5.2 shows that vessels in this cluster mainly travel to and from the Bay of Biscay, through the English Channel heading to Scandinavia by passing through the Skagerrak.

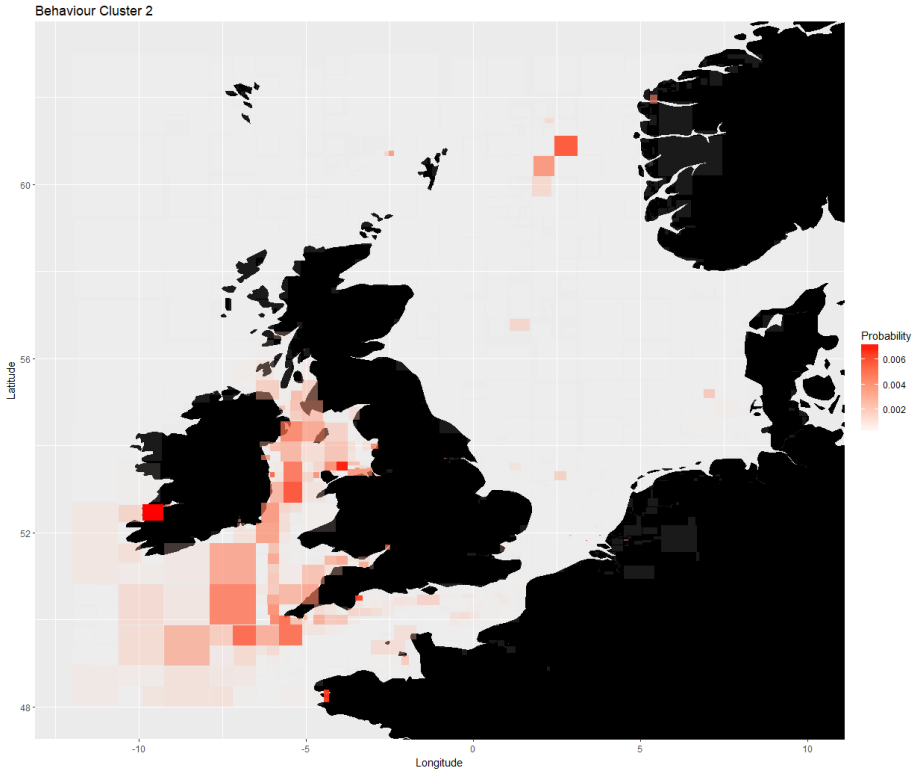


Figure 5.1: The results of applying the Latent Dirichlet Allocation approach for detecting 10 behaviours to a subset of the Global dataset. This is behaviour 2.

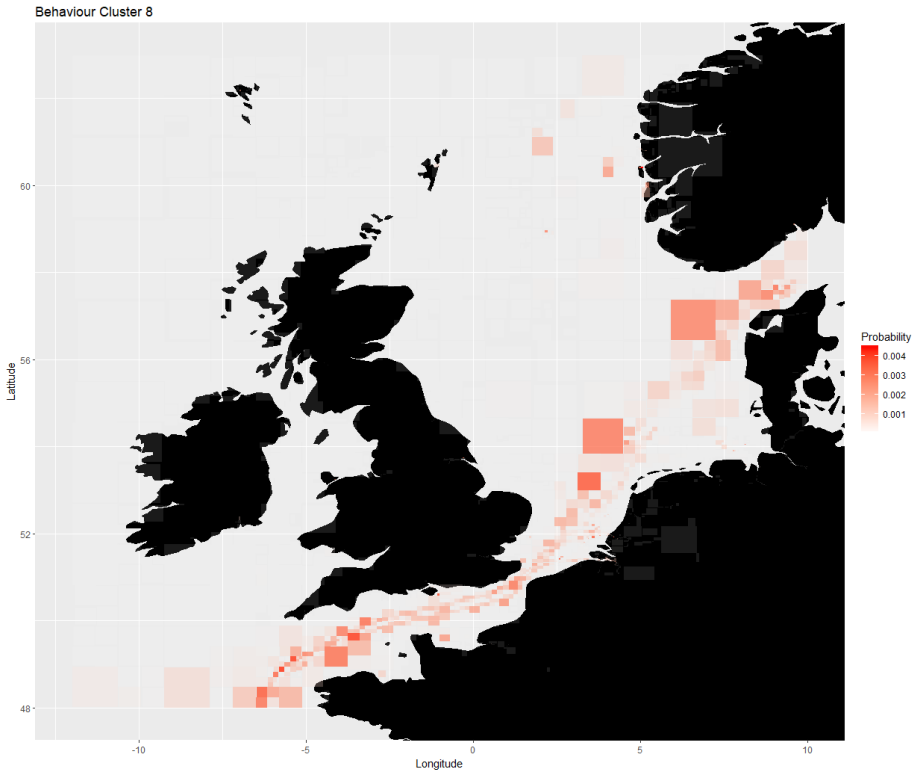


Figure 5.2: The results of applying the Latent Dirichlet Allocation approach for detecting 10 behaviours to a subset of the Global dataset. This is behaviour 8.

5.1.1.1 Mixture of Unigram Anomaly detection for AIS data using symbolic positions

The reported MMSI numbers were chosen to identify ships. Therefore, the data obtained after gridding forms a bag of symbolic geo-locations for each ship (as the example shown in Figure 2.7). This is treated as a document that includes a number of words. A MoU model can be learnt from the set of documents by implementing a Gibbs sampler (described in [145]). Two matrices, the count of document-topic assignments, $\hat{\theta}$, and the count of topic-word assignments, $\hat{\beta}$, (noting that θ and β are the normalised version of $\hat{\theta}$ and $\hat{\beta}$) are recorded. The MoU model provides behavioural patterns of where the ships visit. It can be seen as a number of templates that are generated by analysing the historical way-points. These templates all contribute to deciding whether a ship (i.e., its bag of waypoints) is anomalous or not. Hence, the performance of the anomaly detection is critically dependent

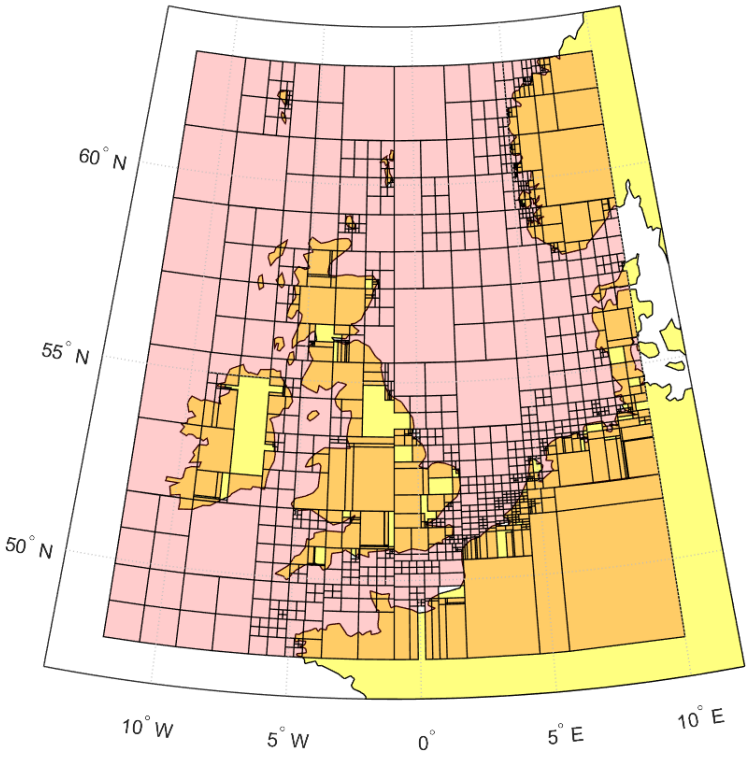


Figure 5.3: The quad-tree adaptive grid used for converting the latitude-longitude coordinates to symbolic representations. Data points within any red rectangles have indices, the rest of the regions are not considered as they did not appear in the training set. Each cell contains less than $Q = 3000$ data points. There are 1183 cells in the grid.

on the robustness and completeness of the clusters. Fortunately, the number of clusters can be estimated in [145] since removing topics is possible within the Gibbs sampler.

The proposed anomaly detection can be performed on any input bag of symbolic positions to calculate an anomaly probability. The likelihood of the ship's waypoints, $p(\boldsymbol{w}|\pi, X, X^0)$, is calculated with the pseudo-counts (recall (2.52) that are identical to the priors, α and η, and the learnt variables $\hat{\theta}$ and $\hat{\beta}$ using (2.47). For $p(\boldsymbol{w}|\pi, X, X^{0'})$, the pseudo-counts were enlarged to construct a model that is based on the training data but makes more use of the prior. The anomaly probability is finally calculated by (2.44).

5.1.1.2 Mixture-of-Unigram Anomaly Detection from Maritime Surveillance Data

The proposed approach was tested on AIS messages from a UK subset of the global dataset (as shown in Figure 5.4). AIS messages are received once per hour, and the number of messages of each track is not fixed, as the AIS transponders can be switched off. The MMSI number of each ship was considered a unique and accurate identity and remove some ships with inconsistent names, types, and invalid latitude-longitude positions. The same test as is described in [85] was considered such that our tests for anomaly detection become several classification problems. The ship data was divided into sets according to five ship types: cargo, tanker, passenger, tug, and other vessels[1]. The models were trained using each set of data and calculate the outlier probabilities for the same sets of data using the models. During the tests, the ships were classified as anomalies if the outlier probability is

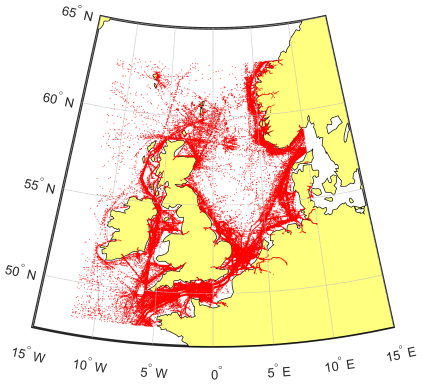
[1]The other vessel class includes fishing ships, towing ships, sailing ships and pleasure craft. Some other types are not chosen simply because there were not enough tracks for training the MoU models.

Model/Data	Cargo	Tanker	Passenger	Tug	Vessel
Cargo	3	375	27	55	193
Tanker	1016	0	39	80	325
Passenger	1528	615	0	110	302
Tug	1718	708	46	0	307
Vessel	1886	739	60	91	0
Total Number of Ships	2456	965	96	275	569

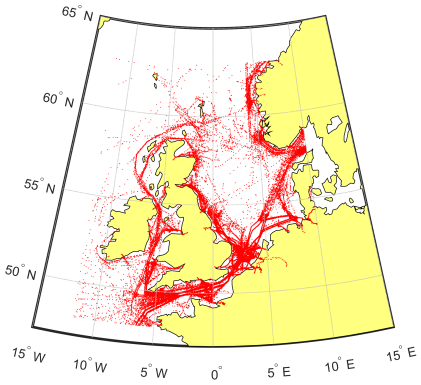
Table 5.1: The number of detected outliers per ship type using the proposed approach with models trained using the AIS data from different ship types (rows) and tested on AIS data from ships of different ship types (columns).

larger than 0.1. The number of anomalies in each set of ships using each model are shown in Table 5.1.

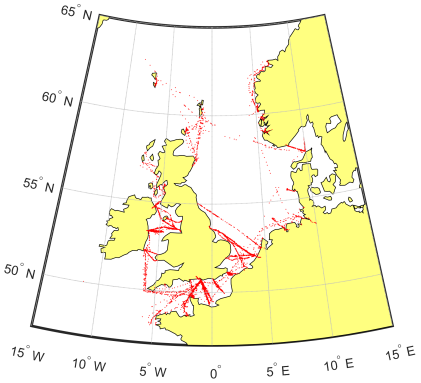
In Table 5.1, the numbers of anomalies yielded by using one of the five models are shown in each row. Each column shows the numbers of anomalies from ship data of the stated type (e.g., when a model was trained on all the tanker data, it detected 39 of the 96 passenger vessels as being anomalous). The last row shows the total number of ships in the set of stated type. By observing the diagonal, the MoU models can be seen to characterise the behaviours well and the anomaly detection algorithm can accurately classify the normal ships. For the rest of the table, as the ships of different types can have similar behaviour (it is possible that a cargo ship's behaviour looks like the behaviour of a tanker ship), it makes sense that some of the ships in one type are classified as normal with respect to another model. Note that the number of ships of each type are imbalanced: for each model, the number of clusters used in the trained MoU models varies. Furthermore, since the number of cargo ships is a lot larger than the other classes, fewer anomalies of other ship types are present when detecting anomalies with respect to the cargo ship model. However, the proposed idea does still detect many anomalies. Some detailed examples of anomalies that are detected by models trained using cargo and tanker ships are shown in Figure 5.5. Each example visualises the difference between the track of the ship and its most probable behaviour that can be detected by the proposed approach. While it is challenging to perform a full quantitative evaluation of performance, this application is perceived to demonstrate the feasibility of the proposed algorithm for detecting anomalies.



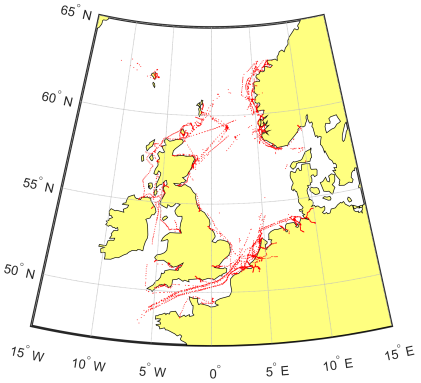
(a) Waypoints of all the cargo ships.



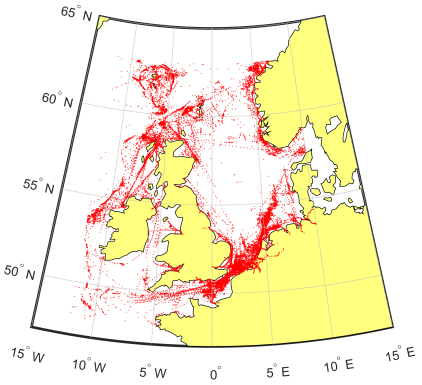
(b) Waypoints of all the tanker ships.



(c) Waypoints of all the passenger ships.

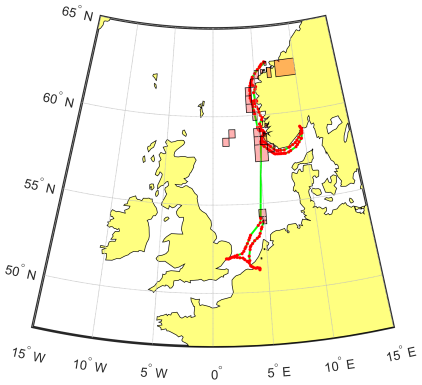


(d) Waypoints of all the tug ships.

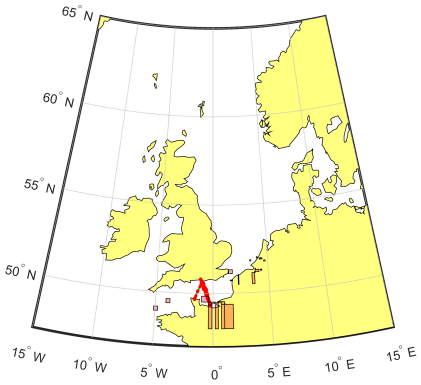


(e) Waypoints of all the other vessels.

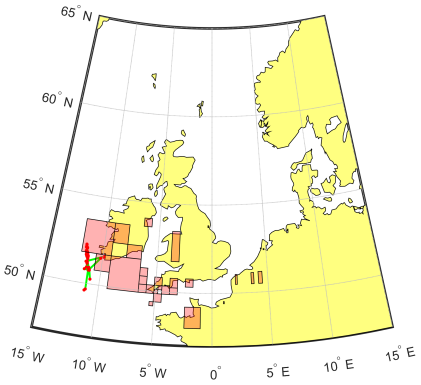
Figure 5.4: The positions of ships that were reported in the AIS messages.



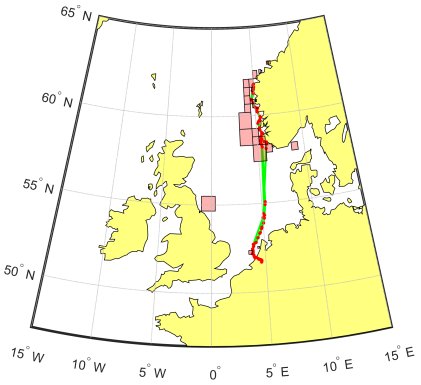
(a) The model is trained using cargo ships, the shown track is a tanker ship which is classified as an anomaly.



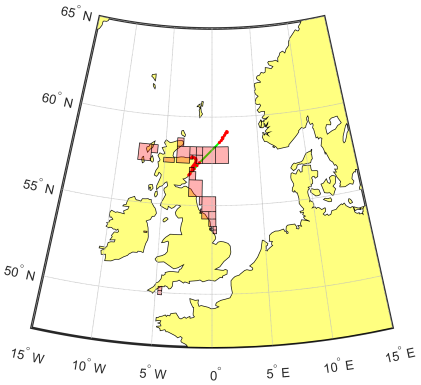
(b) The model is trained using cargo ships, the shown track is a passenger ship which is classified as an anomaly.



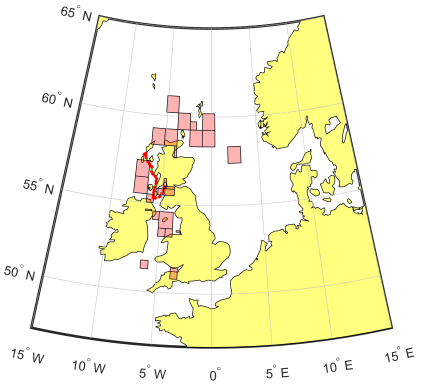
(c) The model is trained using cargo ships, the shown track is an other vessel which is classified as an anomaly.



(d) The model is trained using tanker ships, the shown track is a cargo ship which is classified as an anomaly.



(e) The model is trained using tanker ships, the shown track is a tug ship which is classified as an anomaly.



(f) The model is trained using tanker ships, the shown track is a passenger ship which is classified as an anomaly.

Figure 5.5: Some examples of detected anomalies using models trained with each of a number of types of ship. The most probable behaviour for the trained model is displayed using the red rectangles where the corresponding symbolic geo-positions are most probable. The red points are the way-points and the green lines illustrate the trajectories.

5.2 Change Point Detection

This section uses the change point detection method laid out in section 2.4 and tested against a set of simulated scenarios and then applied to the datasets described in Section 2.8 where the data was abstracted into regions.

5.2.1 Assessment of performance

The change point algorithm is applied to each region independently providing a score of the probability of a change occurring in that region. Ranking these regions provides an operator an order to the areas of interest. To compare the two lists of ranked regions, the true ranks and the algorithmic ranked values, the Spearman's rank correlation coefficient [115] is used.

Spearman's rank is equal to the Pearson's correlation [115] of the rank of a set of vectors. The difference between the two is that Pearson's correlation assesses linear relationships between the two vectors and Spearman's rank correlation assesses monotonic relationships between the two lists of ranks.

For a sample of size n, let X_i and Y_i be the list of scores and R_{X_i} and R_{Y_i} are the scores converted to ranks. Spearman's rank can be expressed as

$$r_s = \frac{cov\left(R_{X_i}, R_{Y_i}\right)}{\sigma_{R_{X_i}} \sigma_{R_{Y_i}}} \tag{5.1}$$

where $cov\left(R_{X_i}, R_{Y_i}\right)$ is the covariance of the rank variables and $\sigma_{(\cdot)}$ represents the standard deviation of the ranks.

Since n, ranks are distinct integers, equation 5.1 can be simplified to

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{5.2}$$

where $d_i = R_{X_i} - R_{Y_i}$.

In the example, laid out in Figures 5.6, 5.7, 5.8 and 5.9, the resultant Spearman's rank correlation coefficient is 1 since the two ranked lists are identical.

The following simulations described in Table 5.2 were run multiple times where the mean Spearman's rank correlation coefficient of the algorithmically generated rank was

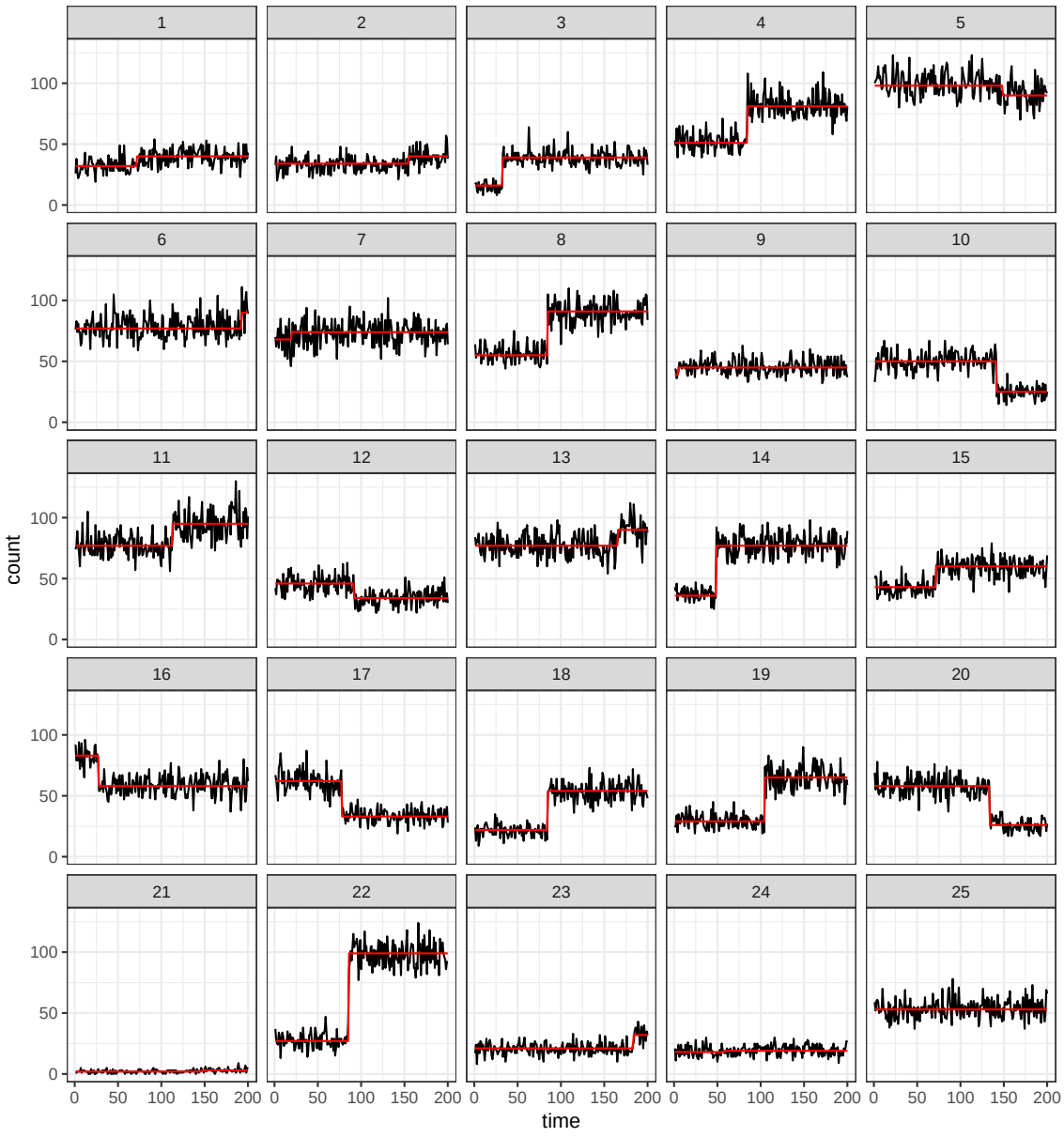


Figure 5.6: 25 simulated regions with change points ordered by geographical region ID, where red denotes the ground truth and black denotes the vessel count. The regions contain time series that remains mostly constant, 2, 21, 24), with some change (3, 12, 15) and large changes (8 22, 14).

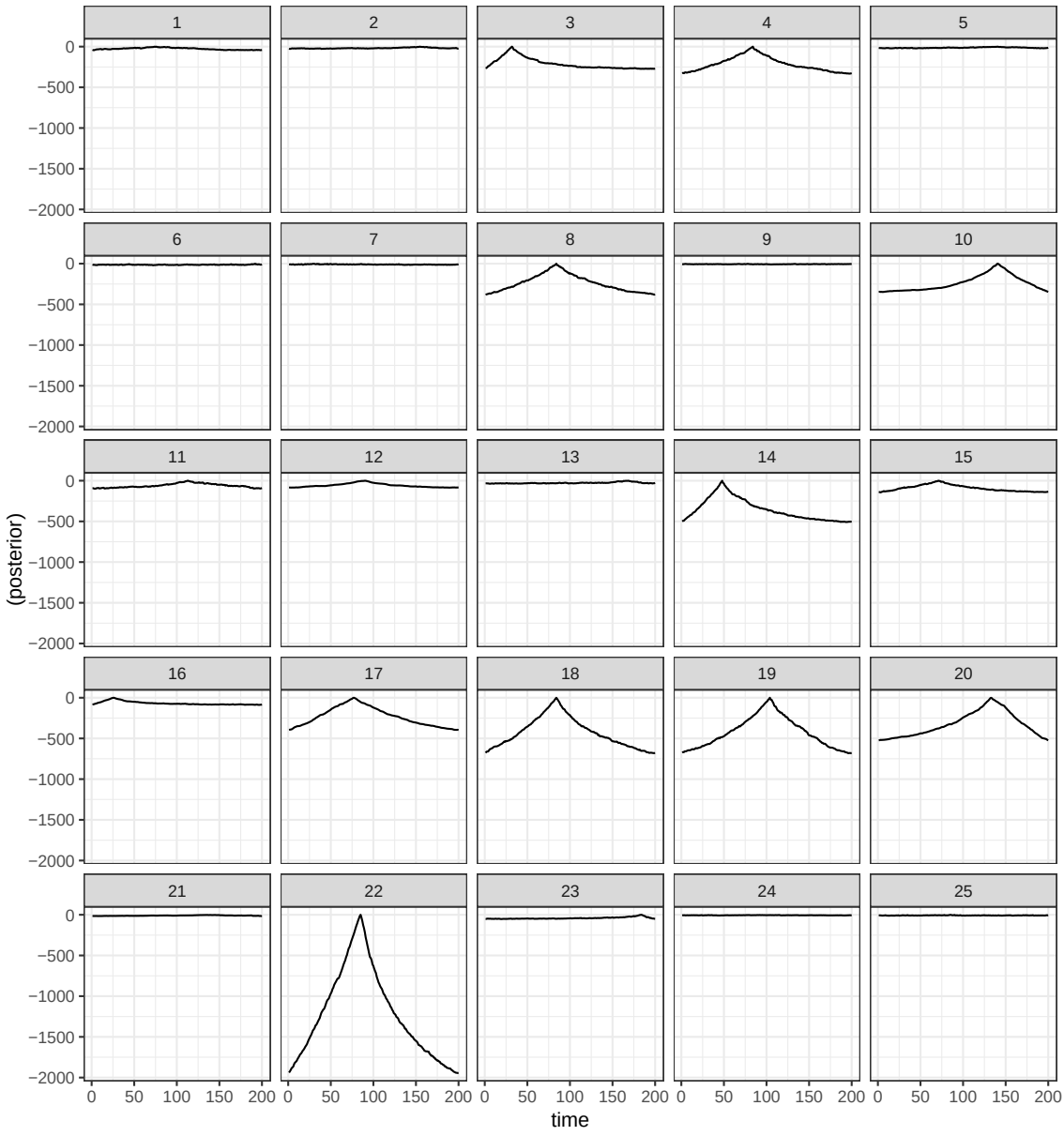


Figure 5.7: 25 simulated regions as seen in Figure 5.6 with the corresponding posterior values plotted.

compared to the mean of a randomly chosen ranked order. Here each scenario had R regions, each with n data in the series with the maximum number of change points in each

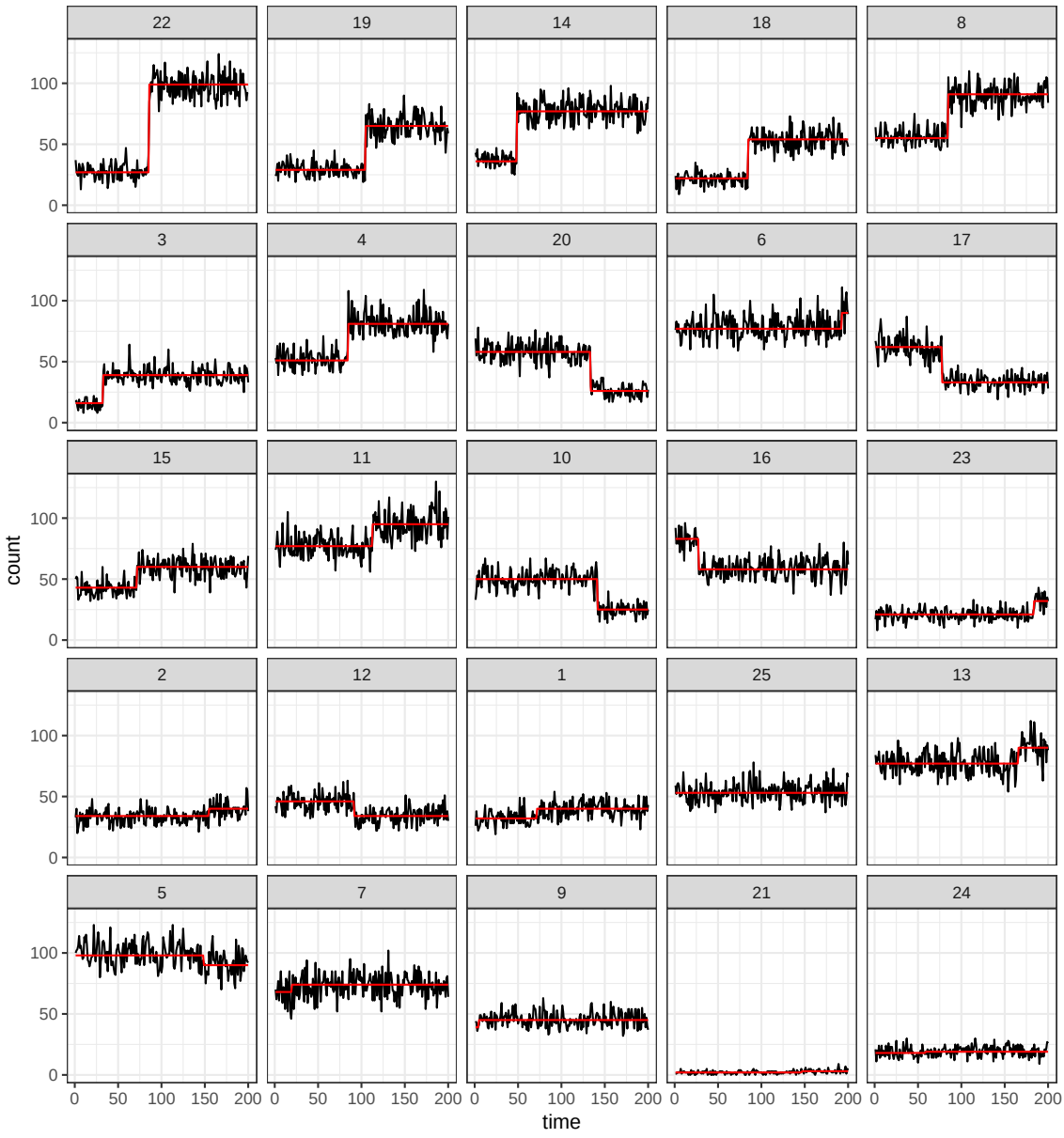


Figure 5.8: The regions presented in Figure 5.6 ordered by the posterior score.

region, c. The mean and standard deviation of the Spearman's rank correlation coefficient was calculated over S simulations. The simulated results show that for all simulations, the random order coefficient r'_s is 0 and all $r_s > 1$. For scenarios with 1 change point, there is a

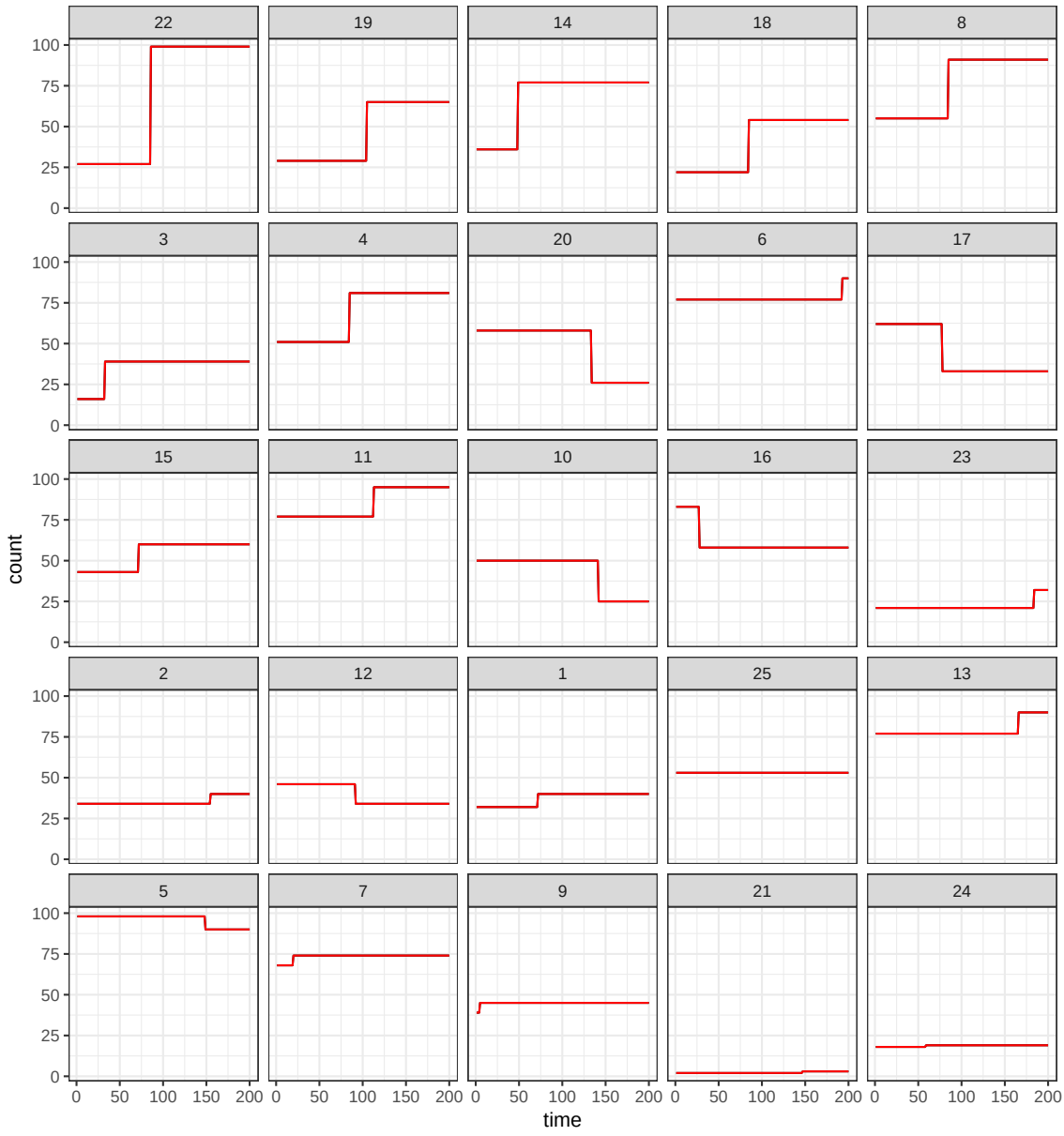


Figure 5.9: The equivalent plot to Figure 5.6 where the ground truth is used to calculate the posterior score.

high correlation between the algorithmic ranked order and ground truth where $r_s > 0.9$. As the number of change points increases the r_s decreases while remaining above 0.

In conclusion, the performance of the change point detection on producing a ranked order based on the size of a change is suitable to provide an operator a list of regions to prioritise effort.

n	R	c	S	r_s		r'_s	
200	25	1	100	0.933	(2.434×10^{-2})	-2.869×10^{-3}	(2.103×10^{-2})
200	25	1	500	0.967	(8.353×10^{-1})	8.848×10^{-4}	(9.942×10^{-3})
500	25	1	500	0.983	(4.343×10^{-2})	1.384×10^{-3}	(8.701×10^{-3})
500	25	10	200	0.873	(1.113×10^{-1})	2.531×10^{-3}	(1.384×10^{-2})
500	25	10	500	0.866	(1.705×10^{-2})	-1.537×10^{-3}	(8.865×10^{-3})
200	100	10	500	0.631	(7.069×10^{-2})	-1.913×10^{-4}	(2.037×10^{-3})
500	100	25	500	0.438	(4.615×10^{-2})	-3.277×10^{-5}	(2.061×10^{-3})
200	1000	25	200	0.263	(7.690×10^{-2})	1.361×10^{-5}	(3.962×10^{-4})
500	1000	25	200	0.326	(2.364×10^{-3})	2.071×10^{-5}	(3.335×10^{-4})

Table 5.2: Results of 10 scenarios each simulated over S runs with R regions, up to c change points, for n length time series. The r_s provides the mean and standard deviation of the Spearman's rank correlation coefficient for the posterior ranking described in Section 2.4 and r'_s denotes the mean and standard deviation of the Spearman's rank correlation coefficient for a random ordered ranking.

5.2.2 Results

The algorithm was tested on the global dataset described in section 2.8.3. Since the AIS messages are received once per hour per vessel, the number of distinct vessels in each region can be counted using the regions generated in section 2.3.1.

The analysis of the global dataset becomes a set of change point tasks.

Figure 5.12 shows the hourly count of MMSI numbers and the log-probability of a change point for the Grid region 1611. The top plot shows the hourly count data. There is a slight hint of a change at 24th January, one day into the dataset, where there is a step down in MMSI count. There is a second peak appearing at 3rd February where the MMSI count begins to increase. The lower plot shows the log-probability of that point is a change point. The peaks at 24th January and 3rd February are evident as are the small peak at 30th January (amongst the low count section) and 5th February (another increase following on from the increase at 3rd February). Figure 5.13 shows the same data from grid region 1611 but this time with a simulated change added to the end of the time series. The top plot shows how the MMSI count data has been augmented by replacing count

data from 2nd February to 6th February (end of time series) with zeros. The lower plot now shows the log-probability of a change point occurring taking into effect the simulated change. The change at 2nd February is successfully detected. Figure 5.10 shows the hourly count of MMSI numbers and the log-probability of a change point for the Grid region 1530. The top plot shows the hourly count data. There is a slight hint of a change at 24th January, one day into the dataset, where there is a step down in MMSI count. There is4nd an increase to a second peak appearing at 3rd February where the MMSI count begins to increase. The lower plot shows the log-probability of that point is a change point. The peaks at 24th January and 3rd February are evident as are the small peak at 1st February. Figure 5.11 shows the same data from grid region 1530 but this time with a simulated change added to the end of the time series. The top plot shows how the MMSI count data has been augmented, again, by replacing count data from 4th February to 6th February (end of time series) with zeros. The lower plot now shows the log-probability of a change point occurring taking into effect the simulated change. The change at 4th February is successfully detected.

The change point algorithm calculates, for each point in the time series, the probability of that point being a change point. This was done for all regions and summarises the change point analysis for each region by calculating the maximum of the posterior probability. From these results the regions were ranked such that there is a higher probability there is a change point in the second plot (simulated change) than the first.

This ties into Dstl's Track Analytics project where they required a method that can determine sensible places to look for changes in vessel counts. The use of the adaptive grid regions and this ranking system allows a region to be prioritised for further inspection and analysis by operators.

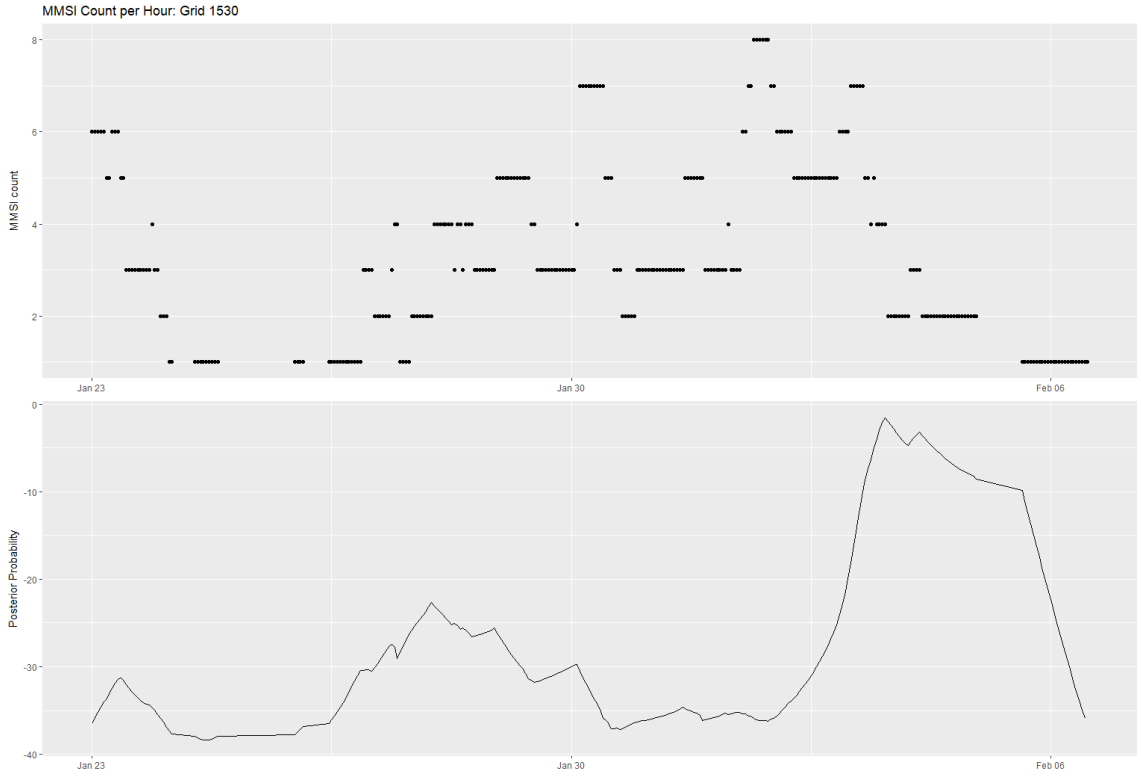


Figure 5.10: The change point method applied to the region with ID 1530 with aggregated MMSI count grouped per hour.

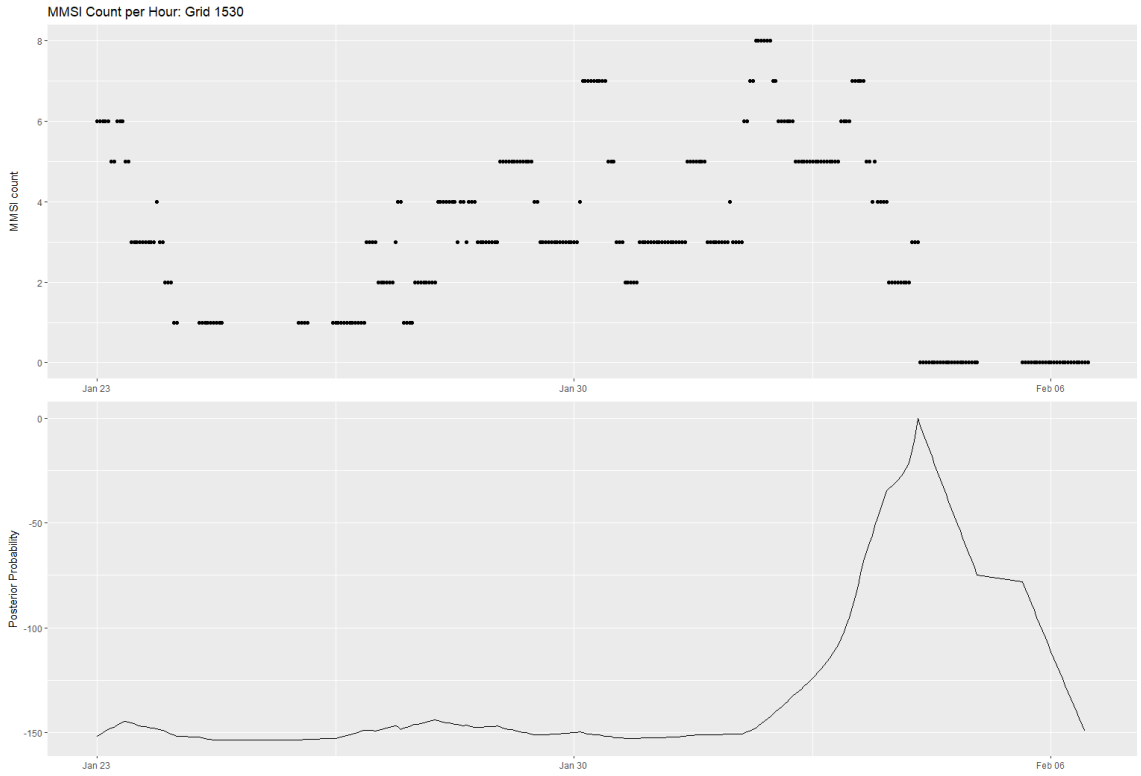


Figure 5.11: The change point method applied to the augmented region with ID 1530 with aggregated MMSI count grouped per hour. Here the augmentation of zeros has been manually added.

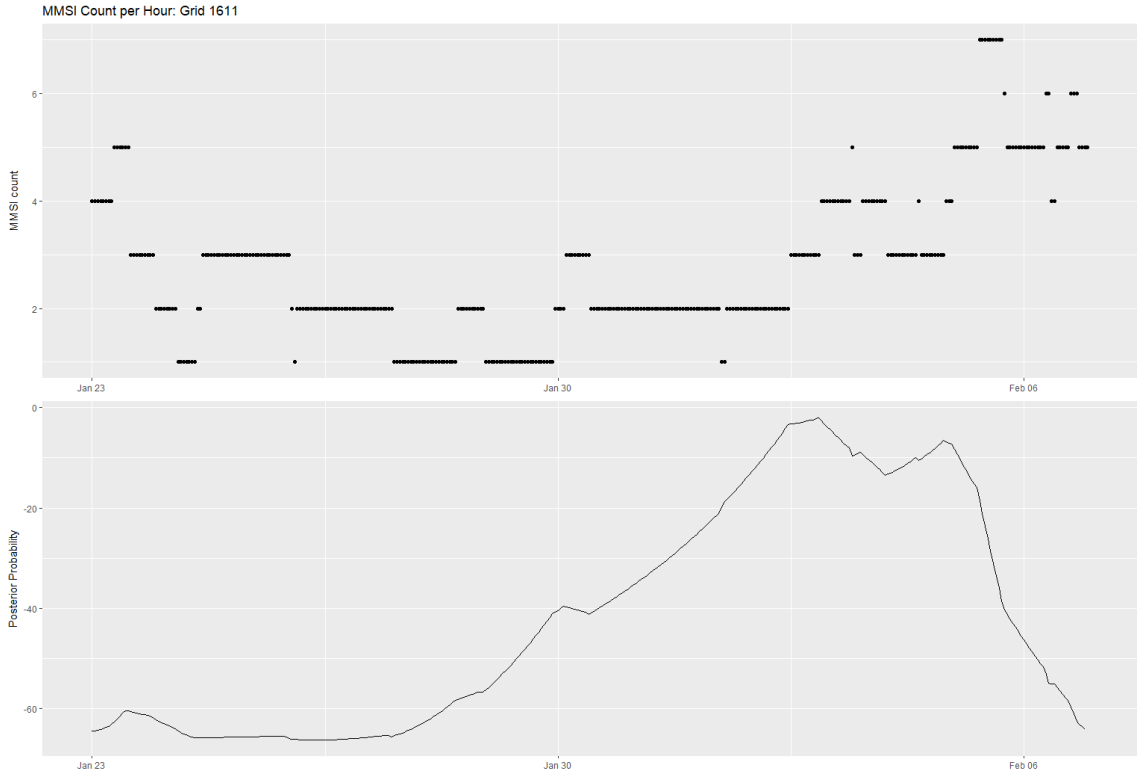


Figure 5.12: The change point method applied to the region with ID 1611 with aggregated MMSI count grouped per hour.

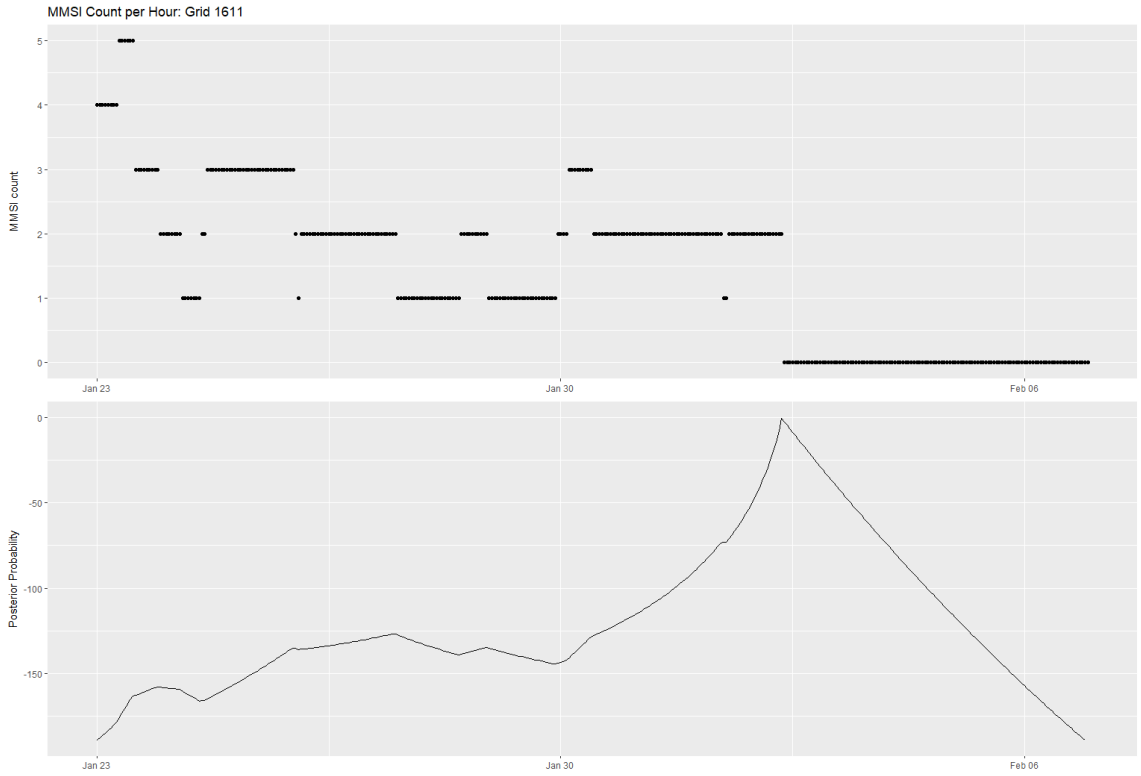


Figure 5.13: The change point method applied to the augmented region with ID 1611 with aggregated MMSI count grouped per hour. Here the augmentation of zeros has been manually added.

5.2.3 Discussion

This section provided a quantification of performance of the change point method described in Section 2.4 by calculating the Spearman's rank correlation coefficient between the rank of regions generated by the algorithm and the ground truth rank of the set of series. This correlation coefficient was compared with the naive geographical order of regions over a set of simulations summarised in Table 5.2. The method was applied to the regions defined by the adaptive grid quad tree in Section 2.3.1.

Chapter 6

Conclusions and Recommendations

6.1 Summary of Thesis Contributions

This thesis has described the process of processing Automatic Identification System (AIS) information to generate a picture of maritime activity and further developed tools to improve our understanding of maritime activities based on vessels behaviour by developing methods to learn individual vessel classes and regional behaviours from the aggregation of AIS data.

Chapter 2 provided an overview of all the methods used throughout the rest of the thesis.

Chapter 3 developed a multiple target tracker (MTT) to produce disambiguated tracklets of vessel movement, assessed the performance of the tracker from a set of simulations using the GOSPA and SIAP metrics, and then applied the tracker to the three AIS datasets. The novelty provided by this chapter is by using a multiple target tracker to disambiguate vessels that are sharing the same MMSI.

Chapter 4 introduced analyses based on single vessels. This took the form of the work presented in Section 4.1 taking the fragmented disambiguated tracklets (where vessel tracks generated by the tracker, within a MMSI, result in several disjoint tracklets) and using track stitching as a post process on the output of the tracker resulting in a singular vessel track. The method of joining tracklets was extended to consider the special case of a ship changing its country of registration, a reflagging event, where a sudden end of a tracklet in

one MMSI coincided with the sudden start of a tracklet in another MMSI. The novelty of this work is by using track stitching to detect the probability of the sudden end of a vessel tracklet on a MMSI to a new tracklet beginning on another MMSI.

Section 4.2 presented a method of using velocity gating on the state estimates of the disambiguated tracks to produce a set of locations where vessels stop. The novelty of this work is detection alternative locations that vessels have stopped, the benefit of this is that this method detects all stopped regions not just those in a port. Chapter 5 introduced analyses based on aggregating data from multiple vessels. This chapter focussed on using a quadtree oriented abstraction of the geography to geospatial regions. Firstly, these symbolic tracks were processed by the text analytics algorithms, using LDA and MoU models, to detect a set of behaviours (Section 5.1.1) and to detect vessel type (Section 5.1.1.1). The novelty of the behavior analysis was in the abstraction of the geospatial data to provide symbolic tracks as documents for text analytics methods. Secondly, the geospatial regions were summarised into hourly count data. Change point detection was applied to these count series to determine the likelihood of a change occurring in the region. The probability of a change point occurring provided a rank for ordering of the set of regions. The novelty of this work focusses on the large number of simultaneous analyses which contrasts from traditional change point detection which focuses on a single time series.

6.2 Recommendations for Future Work

6.2.1 Extensions to Work within the Thesis

Motivated by the work in the thesis, the following extensions could sensibly be implemented;

- Disambiguation can be expanded by changing the transition model options from one (Ornstein-Uhlenbeck) to multiple transition models using an interacting multiple model (IMM) [94] that can be used to model multiple behaviours including moving, turning, and stopped targets.
- Track stitching can be improved by the use of alternative methods such as graph theory [23, 26] or network flow [19, 64] to improve the computational complexity.
- The stopped vessel identification can utilise the model used in the state of the IMM tracker which includes a stationary model in its list of possible models.

- The behaviours for the LDA and MoU methods were based on a quad tree on position, an extension to this would use an adaptive grid that used hyperplanes to split the data over position and velocity to generate the regions. There is a large selection of text analytic models that could be used to infer different characteristics of the symbolic data, for example, the use of a bigram or trigram rather than using a unigram.
- The detection of change points can be developed further to utilise a sliding window to continuously apply the detection algorithm continuously as new count data is received.
- Apply Track Analytics in more applied contexts.

6.2.2 New Directions Motivated by the Thesis

This section provides an additional area of future work which aims to use machine learning algorithms in partnership with human operators. The aim of this study is to prioritise a human operator's actions. This is done by combining the output of a machine learning algorithm, focused on the classifications of events the MLA has made that it is less confident about, and prioritising these events for further inspection by the human operator.

6.2.2.1 Prioritising Targets on the Basis of the Passage of Time and Machine Learning

All chapters preceding this section led to implementations of "Machine Learning Algorithms"[1] on real world cases, for real maritime situational awareness specialists (not "data scientists" - mathematicians/statisticians). By design MLA are trusted by the designers, as they are the people that wrote it. End users on the other hand, have no reason to trust an MLA. Time based scheduling allows a way for end users to be sceptical of an MLA and embrace it rather than have an end user have no trust in an MLA.

There are behaviours of interest, which can be delineated into three sets:

1. B_1, which is the set of behaviours that a machine learning algorithm is trained to recognise (e.g., a cargo vessel behaving in a way that only a cargo vessel can do (e.g., moving in deep water))

[1]Here a Machine Learning Algorithm (MLA) refers to any technique that is being calculated or solved by a system or computer rather than an operator doing analysis "by hand" or manually.

2. B_2, which is the set of behaviours that a machine learning algorithm is not trained to recognise, but a human can spot (e.g., a tug behaving in a way that only a tug can do but which was not in the training set (e.g., moving in open ocean where the pattern of life of the tug type was only present in and around ports))

3. B_3, which is the set of behaviours that are uninteresting (e.g., a fishing vessel in a known fishing area). This class also incorporates vessels acting against their ship type (e.g., a fishing vessel pretending to be a cargo vessel to fish in an area it is not allowed to fish in.)).

Behaviours transition to B_1 and B_2 from B_3 with known relatively slow rates. Vessels cannot transition back to B_3. Once a vessel has displayed a behaviour, i.e., there is evidence that you are a cargo vessel, the vessel is assumed that it will always be a cargo (i.e., B_3 describes objects that truly are indistinguishable from the normal behaviour of that vessel type).

For an unknown current behaviour, B_t, it is assumed that

$$p(B_t \in B_1 | B_t \in B_1 \cup B_2), \tag{6.1}$$

is known, i.e., the probability that the machine learning algorithm saw the behaviour in the training set given that it was interesting. This is a parameter of the system. An object can be observed over time which will produce the following estimate $P(\cdot)$.

An object is observed over time and the machine learning algorithm provides a current estimate of

$$p(B_t \in B_1 | y_{1:t}, B_t \in B_1 \cup B_3). \tag{6.2}$$

When an operator looks at an object, they tell you whether

$$B_t \in B_1 \cup B_2 \tag{6.3}$$

(if B_t is interesting) with "100%" accuracy. We can observe an operator looking at an object and calculate the time since an operator had last looked at an object. The time since the person last looked is Δ.

The question is then what is

$$p(B_t \in [B_1 \cup B_2] | y_{1:t})? \tag{6.4}$$

This is an interpolation between a function derived from Δ and the output from the machine learning algorithm, i.e.,

$$p(B_t \in B_1 | y_{1:t}, B_t \in [B_1 \cup B_3]). \tag{6.5}$$

The states are described as follows;

- The machine learning algorithm:

$$P(B_1 | y_{1:n}, t) = 1 - P(B_2 \cup B_3 | y_{1:n}) \tag{6.6}$$

- The Human:

$$P(B_1 \cup B_2 | y_{1:n}, t) = \sum_i P(B_i | t) = P_H[1 - \text{Po}(0; \lambda_i t)] \tag{6.7}$$

  where P_H is the fallibility of a human, $\text{Po}(0; \lambda_i t)$ is the probability of nothing happening of interest and the complement of probability of nothing happening is the probability of something interesting has happened.

- The probability of something interesting:

$$= \frac{P(B_1 | y_{1:n}, t)}{\bigcup_i P(B_i | y_{1:n}, t)} \tag{6.8}$$

$$= \frac{P(B_1 | y_{1:n}, t)}{P(B_1 | y_{1:n}, t) + P(B_2 | y_{1:n}, t) + P(B_3 | y_{1:n}, t)} \tag{6.9}$$

$$= \frac{1 - P(B_2 \cup B_3 | y_{1:n})}{1 - P(B_2 \cup B_3 | y_{1:n}) + P_H[1 - \text{Po}(0; \lambda_i t)] + [1 - P(B_1 \cup B_2 | y_{1:n}, t)]} \tag{6.10}$$

$$= \frac{1 - P(B_2 \cup B_3 | y_{1:n})}{1 - P(B_2 \cup B_3 | y_{1:n}) + P_H[1 - \text{Po}(0; \lambda_i t)] + [1 - P_H[1 - \text{Po}(0; \lambda_i t)]]} \tag{6.11}$$

The important assumption needed to be defined, and to be recognised, is that a machine learning algorithm's training set is **finite** and that there is a probability that there is an **event of interest** that is not in the finite training set. The next part of this assumption

is that the event of interest is in the human operator's training set and the operator has some ability to detect it and can unambiguously determine if the object in the event was behaving in a fashion of interest.

Following on from these assumptions, the ability of the machine learning algorithm can be quantified. Assuming that 1 in a 100 events is not in a machine learning algorithm's training set, the only way to see events of this type is if a human operator looks at the object and makes a judgement as to whether the event has happened or not. There is no machine learning algorithm that can match the operator's abilities and only the operator will be able to tell if this event has occurred.

From this, two questions can be defined;

1. *Has the object behaved in a particular fashion that a machine learning algorithm has seen before (i.e., in its training set)?*

2. *How long has it been since the operator looked at the object?*

These two questions together inform the probability that the object is behaving in a way that is of interest to the operator. The first is the output from the MLA while the second acts as a time-since-last-looked module.

The extreme examples of the problem can be defined when

1. the probability that the event is in the training set is 1, then the system behaves like a normal machine learning algorithm with no input from the time-since-last-looked module.

2. if the probability that the event is in the training set is 0, then the system runs the time-since-last-looked module indicating the targets that have not been viewed for certain lengths of time.

When the probability of the event is in the training set is in the interval $(0, 1)$, the system provides an interpolation between the output of the machine learning algorithm and what the passage of time dictates.

Given that there is a machine learning algorithm which can be assumed has the ability to detect behaviours reliably for behaviours in its training set but not necessarily all behaviours, and given that the chance that a behaviour happens or has happened increases as time evolves, the aim is to, for a given object the operator last saw at time t and a particular

output from the machine learning algorithm, calculate the probability that the object is behaving in a fashion that is of interest.

Traditionally, this problem would rest wholly on the machine learning algorithm by prioritising operator activity on the basis of what a machine learning algorithm outputs. The key of interpreting this scenario is to not prioritise operator activity based on the output from the machine learning algorithm but, rather, to prioritise operator activity based on the output of the machine learning algorithm *and* factoring in that the machine learning algorithm is known not to be omniscient (i.e., not all behaviours are in the training set).

The important distinction is not whether or not the machine learning algorithm can detect it but if an operator would detect it if the operator looks (but an operator has to look to be able to detect). The longer the time elapsed since an operator last observed an object, the greater the chance that an event of interest has occurred, so the probability that an operator would detect it also increases because the probability that it increases is based on the evidence that some behaviour is happening at a point in time and it is whether or not an operator looked determines whether or not the behaviour is detected.

This can be applied to each of a large number of objects and prioritisation of targets on the basis of the combined output rather than only one of the two methods; machine learning algorithm and human operator. This combined output was measured to determine whether or not it makes a difference to an operator's ability to detect objects behaving in an interesting fashion. The important element is to recognise that the machine learning algorithm's training set is finite and that there is a probability that the event of interest was never in the training set. The assumption made is that the event is in the operator's training set and the operator has some ability to detect it and can unambiguously determine if the object in the event was behaving in a fashion of interest.

6.2.2.1.1 Survival Analysis and Censored Data

Censoring problems arise when not all the data are directly observed. For this set of problems, a censored likelihood needs to be used [73]. This section is intended to outline the principles of censored likelihoods with applications for calculating likelihoods of incomplete observations and then to present a method of iteration that would translate into the collaborative learning method [16].

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with

probability density function

$$f(x|\theta) \tag{6.12}$$

Each variable is censored if it exceeds C_i, a known constant and can be expressed as $Pr(X_i \geq C_i|\theta)$ such that

$$Pr(X_i \geq C_i) = \int_{C_i}^{\infty} f(x_i|\theta)dx \tag{6.13}$$

$$= \Big[F(x_i|\theta)\Big]_{C_i}^{\infty} \tag{6.14}$$

$$= -F(C_i|\theta) \tag{6.15}$$

where

$$F(x) = \int f(x)dx \tag{6.16}$$

Thus, the censored likelihood can be expressed as

$$\prod_{i=1}^{n} f_X(x_i|\theta)^{\delta_i} Pr(X_i \geq C_i|\theta)^{1-\delta_i}. \tag{6.17}$$

where δ_i is the censored observation indicator following the following distribution,

$$\delta_i = \begin{cases} 1 & \text{if event of interest observed,} \\ 0 & \text{if event of interest censored.} \end{cases} \tag{6.18}$$

and $Pr(X_i \geq C_i|\theta)$ is the censored distribution.

The aim of using this technique is in situations where the set of observations of a state, for example, assessing the lifetime of a patient after treatment [65, 107, 120], estimating health care costs from incomplete information [81, 143], time to failure in manufacturing [111] and the time to a vessel behaviour being of interest ($B_3 \rightarrow B_2$) where the MLA has no training information. There are two possible types of observation that can be encountered.

1. Direct observations of times vessels exhibit behaviours of interest,

2. Incomplete information on any change to the behaviour of interest.

$f(x|\theta)$ is the distribution of the time to behaviours of interest since the last observation and $Pr(X_i \geq C_i|\theta)$ represents the contribution of a vessel that has a behaviour that is not of interest at the current observation C_i.

$Pr(X_i \geq C_i|\theta)$ is the probability that the vessel will change its behaviour to one of interest sometime from the latest observed time to ∞.

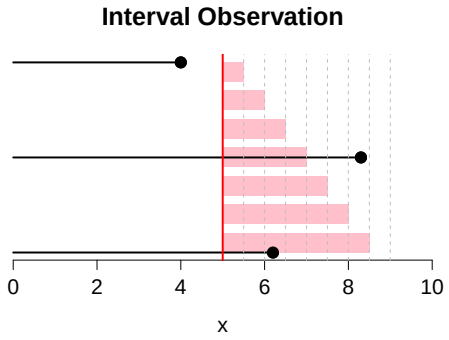


Figure 6.1: Visualisation of time since last looked.

Combining the theory above with that of censored data and survival analysis, it can be expressed in terms of observations of the object by an operator as censored thresholds. Consider an object that has a given behaviour and at a certain point in time t it changes behaviour to be of interest. An object very rarely will transition to the interesting behaviour. This has been visualised in Figure 6.1, where the targets are represented as black lines and the event of interest is the black circle at the right of the line. The observer last looked at the target at the vertical red line. As time passes after the observed time (pink segments in Figure 6.1), the probability that the event of interest for a target occurs in the interval of observed time and current time can be calculated.

Observations can only be taken at discrete time-steps, e.g., seconds. An observation can be separated by several units of time. An observation is only recorded if the object is actually looked at. If it has changed behaviour the likelihood would be distributed by a given probability density function, $f(x_i|\theta)$. If the behaviour has not changed to interesting, i.e., the behaviour remains uninteresting, i.e., still 0, then $Pr(X_i \geq C_i|\theta)$ is the likelihood describing the distribution of the object changing behaviour sometime in the future, i.e., in the interval (C_i, ∞), where C_i is the threshold at time i (last seen).

This time-since-last-looked requires a slightly different approach to the censoring problem

since it is motivated by the need to know the probability of when an object might change behaviour rather than the probability an object would ever change. So, the required probability is a motivation for an operator to look at the object.

The resulting calculation for the probability of a censored object is the hazard function, which equates to the integral of the distribution function over the interval (C_i, ∞). Using the Riemann sum approximation to the integral, this can be redefined as a recursive function. So, if an object is observed in state 0 at time $t = C_i$, as time goes by, and you are not looking at the object, a probability that the object would have changed state if you looked **now**, D_i (this can be rephrased as *since you last looked*) would be defined as

$$Pr(D_i \geq X_i \geq C_i) = \int_{C_i}^{D_i} f(x_i|\theta)dx \tag{6.19}$$

$$= \Big[F(x_i|\theta)\Big]_{C_i}^{D_i} \tag{6.20}$$

$$= F(D_i|\theta) - F(C_i|\theta) \tag{6.21}$$

where C_i is the current observation (at a time in the past), and D_i is the next potential observation (the current time).

The probability at $t + \tau$, where $t + \tau > t$, of the change of state has occurred in the interval $(t, t + \tau]$ is

$$Pr(t + \tau \geq X_i \geq t) = \int_{t}^{t+\tau} f(x_i|\theta)dx \tag{6.22}$$

$$\tag{6.23}$$

Using the Riemann sum approximation, it becomes

$$Pr(t + \tau \geq X_i \geq t) = \int_{t}^{t+\tau} f(x_i|\theta)dx \tag{6.24}$$

$$\approx \sum_{\delta=t}^{t+\tau} f(x_\delta|\theta) \tag{6.25}$$

$$\equiv \mathcal{F}_{(t+\tau,t)} \tag{6.26}$$

As time increases the formula above can be converted to an iterative process as such

$$Pr(t+\tau+1 \geq X_i \geq t) = \int_{t}^{t+\tau+1} f(x_i|\theta)dx \tag{6.27}$$

$$\approx \sum_{\delta=t}^{t+\tau+1} f(x_\delta|\theta) \tag{6.28}$$

$$= \mathcal{F}_{(t+\tau,t)} + f(x_{t+\tau+1}|\theta) \tag{6.29}$$

$$\tag{6.30}$$

The novelty of this work is explored through the calculation of equation 6.21 from the observed time to the current time rather than the traditional method using equation 6.13. The combination of the machine learning algorithm output and this passage of time since last observed will be continued in future work. Alongside this new scheduling algorithm, the developed process performance of the machine learning algorithm will be optimised such that it will guarantee that the scheduling algorithm does not degrade the performance of that of the machine learning algorithm on its own, i.e., the scheduling algorithm only improves the bad answers of the machine learning algorithm.

6.2.2.2 Closing Remarks

This would represent a significant extension of the work of the thesis and represents an exciting avenue for new research.

Appendix A

Collaborations

Several government projects were closely linked with this thesis. These include: the Dstl Track Analytics project, the Dstl Stone Soup project, and the application of these projects to the Royal Navy Project NELSON and Information Warrior, Hackathon and Codeathons.

A.1 Track Analytics

The Dstl Track Analytics project was awarded to the University of Liverpool in 2018 and is coupled significantly with the theme of this PhD. The aim of this project is to develop a working environment for tracking, detecting anomalies and classifying vessels from their behaviour [91]. This links to the adaptive grid work and the associated behaviour analysis utilising the text analytic models of Mixture-of-Unigrams [103] and latent Dirichlet allocation [11] in Chapter 5. The Track Analytics project has six tasks;

Task 1 Tracking. Track and fuse AIS ship position reports and satellite data to generate a track of a vessel's history.

Task 2 Classification. From the ships movement determine the class (type) of vessel. Some vessels broadcast their type however many don't. This technique can be used to give a level of confidence in the information provided by the vessel.

Task 3 Clustering. There will be some ships similar to other ships, with the aim of being able to identify them.

Task 4 Anomaly detection. Looking for vessels exhibiting behaviours inconsistent with their class or normal activity

Task 5 Search. Search the database using complex queries.

Task 6 Prediction. Use of a trained neural network to determine the most likely destination of a vessel based on its previous passages and its location.

The intent is to get the software used by the UK government, potentially through the NMIC.

The current implementation comprises of a Mongo database [25, 82] distributed over several computers and the analysis is done in a central single computer.

It is planned to next test the applicability of the algorithms to the land domain in an urban setting. For example, an urban scenario in which a camera is mounted to a helicopter looking down at a city. The model tests the ability of the model to track all the vehicles, people, and do classification, clustering, anomaly spotting, search, and machine learning on vehicles moving through the city. All achieved through the simulation of both an urban environment and using an imaginary helicopter looking at an imaginary city. Using the model, a user can determine the benefits of changing the sensor suite, for example a higher resolution camera, adding satellites to the mix or possibly re-siting the sensors.

A.2 Collaboration with Dstl and the National Maritime Information Centre

Alongside the work carried out as part of the Track Analytics project, a collaboration with Stephen Ablett (Dstl and National Maritime Information Centre) has provided a detailed knowledge of the domain and examples of the existing systems used by the NMIC for visualising maritime information.

The Defence Science and Technology laboratory (Dstl) has been investigating techniques to improve MSA for at least 10 years [75, 74]. This activity has supported the development of now operational capabilities and continues to support the development of tools and techniques.

This collaboration took the form of Stephen Ablett visiting the University of Liverpool on several occasions I visited the NMIC (Portsmouth) to look at government data and meet the team working there.

Much of the time spent at the NMIC provided excellent awareness of the need for the output of the Track Analytics project and by extension the contents of this thesis.

A.3 Stone Soup

The Defence Science and Technology Laboratory (Dstl) teamed up with the University of Liverpool to build an open source tracking and state estimation toolkit that also benefits support from international collaboration both within academia and industry [128, 129]. As a result, Stone Soup was launched at FUSION 2019 with a beta release.

A.4 Project NELSON

NELSON is an innovation programme within the Royal Navy. It is focussed on using artificial intelligence and data science to build a "Ship's Mind", enabling better decision making [28].

Project NELSON has set up a development environment enabling access to different streams of information using a GraphQL API [43, 126, 44] to support the testing and integration of new tools.

Participation in a number of Hackathons has provided excellent opportunities to develop and test some of the algorithms developed as part of this PhD with real data in a live scenario (which was able to utilise both the Track Analytics project and the Stone Soup framework).

Appendix B

AIS Specifications

B.1 Message Types

Table B.1: AIS Message Types

Message Type	Description
01	Position Report Class A
02	Position Report Class A (Assigned schedule)
03	Position Report Class A (Response to interrogation)
04	Base Station Report
05	Static and Voyage Related Data
06	Binary Addressed Message
07	Binary Acknowledge
08	Binary Broadcast Message
09	Standard SAR Aircraft Position Report
10	UTC and Date Inquiry
11	UTC and Date Response
12	Addressed Safety Related Message
13	Safety Related Acknowledgement
14	Safety Related Broadcast Message
15	Interrogation
16	Assignment Mode Command
17	DGNSS Binary Broadcast Message

... continued

Message Type	Description
18	Standard Class B CS Position Report
19	Extended Class B Equipment Position Report
20	Data Link Management
21	Aid-to-Navigation Report
22	Channel Management
23	Group Assignment Command
24	Static Data Report
25	Single Slot Binary Message
26	Multiple Slot Binary Message With Communications State
27	Position Report For Long-Range Applications

B.2 Message Variables

Table B.2: AIS Message Payload Variables

Information Item	Type	Information generation, type and quality of information
MMSI	Static	Set on installation
Call Sign and Name	Static	Set on installation
IMO Number	Static	Set on installation
Length and Beam	Static	Set on the installation of the AIS equipment as it is the distance from the position of the AIS transceiver to the port, starboard, bow and stern of the vessel that are then combined to calculate the length and width)
Type of Ship	Static	Selected from a pre-installed list. See Table B.5..
Location of position fixing antenna	Static	Set on installation (see also comment on length and beam)

. . . continued

Information Item	Type	Information generation, type and quality of information
Ship position (with accuracy indication and integrity status)	Dynamic	Automatically updated from the onboard GPS equipment connected to the AIS equipment. The accuracy indicator classifies either if the signal has better than 10m accuracy or worse than 10m accuracy.
Position Time Stamp in UTC	Dynamic	Automatically updated from the GPS equipment. (Not included in all transmitted message types
Course over Ground (COG)	Dynamic	Automatically updated from the GPS equipment. Some GPS equipment cannot calculate the COG and thus the information might not be available.
Speed over Ground (SOG)	Dynamic	Automatically updated from the GPS equipment. Some GPS equipment cannot calculate the SOG and thus the information might not be available.
Heading	Dynamic	Automatically updated from the vessel's heading sensor connected to the AIS equipment.

. . . continued

Information Item	Type	Information generation, type and quality of information
Navigational Status	Dynamic	The ship's Officer of the Watch (OOW) is required to update the navigational status as necessary. • Underway by engine • at anchor • not under command (NUC) • restricted in ability to manoeuvre (RIATM) • moored • constrained by draught • aground • engaged in fishing • underway by sail These all conform to the International Regulations for Preventing Collisions at Sea.
Rate of Turn (ROT)	Dynamic	Automatically updated from the rate of turn sensor or derived from the ship's gyro. Some ships will not have ROT or gyros connected to their AIS equipment and as such the information might not be available.
Ship Draught	Dynamic	The draught is manually entered at the start of each voyage (defined as from one port to another port). The maximum draught is entered at the start of the voyage and updated as required throughout the voyage, e.g. as a result of debiasing before entering a port.

. . . continued

Information Item	Type	Information generation, type and quality of information
Hazardous Cargo	Dynamic	The Hazardous Cargo Type is entered at the start of a voyage and informs whether or not hazardous materials are being transported. This only indicates if a hazardous material is being carried and not the quantity. The options are: • Dangerous Goods (DG) • Harmful substances (HS) • Marine Pollutants (MP)
Destination and Estimated Time of Arrival (ETA)	Dynamic	Manually entered at the start of a voyage. This will be updated as necessary throughout the voyage (particularly the ETA, if issues arise during the voyage). A manually entered port name also falls foul of spelling mistakes amongst other things, e.g. destinations such as “Hell”.

B.3 Navigation Status

Value	Description
0	Under way using engine
1	At anchor
2	Not under command
3	Restricted manoeuvrability
4	Constrained by draught
5	Moored
6	Aground
7	Engaged in Fishing
8	Under way sailing
9	Reserved for future amendment of Navigational Status for HSC
10	Reserved for future amendment of Navigational Status for WIG
11	Power-driven vessel towing astern (regional use)
12	Power-driven vessel pushing ahead or towing alongside (regional use)
13	Reserved for future use
14	AIS-SART is active
15	Not defined (default)

Table B.3: Navigation Status

B.4 Rate of Turn

Value	Description
−127	turning left at more than 5deg per 30 seconds
−126 - 1	turning left at up to 708deg per minute or higher
0	Not Turning
1 - +126	turning right at up to 708deg per minute or higher
+127	turning right at more than 5deg per 30 seconds

Table B.4: Rate of Turn

B.5 Ship Type

Table B.5: Ship Types

Code	Ship Type
00	Ship not providing Ship Type
01 - 19	Reserved for future use
20	Wing In Ground - (ALL)
21	Wing In Ground - Carrying DG, HS, or MP, IMO hazard or pollutant category A
22	Wing In Ground - Carrying DG, HS, or MP, IMO hazard or pollutant category B
23	Wing In Ground - Carrying DG, HS, or MP, IMO hazard or pollutant category C
24	Wing In Ground - Carrying DG, HS, or MP, IMO hazard or pollutant category D
25 - 29	Wing In Ground - Reserved for future use
30	Fishing vessels
31	Towing vessels
32	Towing and length of the tow exceeds 200m or breadth exceeds 25m
33	Vessels engaged in dredging or underwater operations
34	Vessels engaged in diving operations
35	Vessels engaged in military operations
36	Sailing vessels
37	Pleasure craft
38 - 39	Vessel - Reserved for future use
40	High Speed Craft - (ALL)
41	High Speed Craft - Carrying DG, HS, or MP, IMO hazard or pollutant category A
42	High Speed Craft - Carrying DG, HS, or MP, IMO hazard or pollutant category B
43	High Speed Craft - Carrying DG, HS, or MP, IMO hazard or pollutant category C

. . . continued

Code	Ship Type
44	High Speed Craft - Carrying DG, HS, or MP, IMO hazard or pollutant category D
45 - 48	High Speed Craft - Reserved for future use
49	High Speed Craft - No additional information
50	Pilot Vessel
51	Search and Rescue vessels
52	Tugs
53	Port tenders
54	Vessels with anti-pollution facilities or equipment
55	Law enforcement vessels
56	Spare – for assignments to local vessels
57	Spare – for assignments to local vessels
58	Medical transports (as defined in the 1949 Geneva Conventions and Additional Protocols)
59	Ships and aircraft of States not parties to an armed conflict (Noncombatant ship according to RR Resolution No. 18)
60	Passenger Ships - (ALL)
61	Passenger Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category A
62	Passenger Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category B
63	Passenger Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category C
64	Passenger Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category D
65 - 68	Passenger Ships - Reserved for future use
69	Passenger Ships - No additional information
70	Cargo Ships - (ALL)
71	Cargo Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category A

. . . continued

Code	Ship Type
72	Cargo Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category B
73	Cargo Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category C
74	Cargo Ships - Carrying DG, HS, or MP, IMO hazard or pollutant category D
75 - 78	Cargo Ships - Reserved for future use
79	Cargo Ships - No additional information
80	Tanker - (ALL)
81	Tanker - Carrying DG, HS, or MP, IMO hazard or pollutant category A
82	Tanker - Carrying DG, HS, or MP, IMO hazard or pollutant category B
83	Tanker - Carrying DG, HS, or MP, IMO hazard or pollutant category C
84	Tanker - Carrying DG, HS, or MP, IMO hazard or pollutant category D
85 - 88	Tanker - Reserved for future use
89	Tanker - No additional information
90	Other Types of Ship - (ALL)
91	Other Types of Ship - Carrying DG, HS, or MP, IMO hazard or pollutant category A
92	Other Types of Ship - Carrying DG, HS, or MP, IMO hazard or pollutant category B
93	Other Types of Ship - Carrying DG, HS, or MP, IMO hazard or pollutant category C
94	Other Types of Ship - Carrying DG, HS, or MP, IMO hazard or pollutant category D
95 - 98	Other Types of Ship - Reserved for future use
99	Other Types of Ship - No additional information

Table B.5: Ship Types

Appendix C

General Information

C.1 List of Acronyms

Table C.1: Acronyms

Acronym	Description
AIS	Automatic Identification System
AIS-SART	AIS Search and Rescue Transmitter
ARPA	Automatic Radar Plotting Aids
AtoN	Aids to Navigation
CMTS	Committee of Maritime Transportation System
COS	Course Over Ground
DAC	Digital to Audio Converter
DG	Dangerous goods
EEZ	Exclusive Economic Zone
EKF	Extended Kalman Filter
ETA	Estimated Time of Arrival
FATDMA	Fixed Access Time Division Multiple Access
FOV	Field of View
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HS	Harmful substances
HSC	High-Speed Craft

. . . continued

Acronym	Description
IMM	Interacting Multiple Model
GOSPA	Generalised Optimal Sup-Pattern Assignment
IALA	International Association of Marine Aids to Navigation and Lighthouse Authorities
IMO	International Maritime Organisation
ITDMA	Incremental Time Division Multiple Access
ITU	International Telecommunication Union
JPDAF	Joint Probabilistic data association filter
KF	Kalman Filter
LDA	Latent Dirichlet Association
LEO	Low Earth Orbit
LES	Land Earth Station
LRIT	Long-Range Identification and Tracking
MDA	Maritime Domain Awareness
MEO	Medium Earth Orbit
MID	Maritime identification digits
MLA	Machine learning algorithm
MMSE	Minimum mean-squared error
MMSI	Maritime Mobile Service Identity
MoU	Mixture of Unigrams
MSA	Maritime situational awareness
MP	Marine pollutants
MSC	Maritime Safety Committee
MTT	Multiple target tracker
MWAS	Maritime Wide Area Surveillance
NM	Nautical Miles
NMIC	National Maritime Information Centre
NMEA	National Marine Electronics Association
OCINF	Oil Companies International Marine Forum
OSPA	Optimal Sup-Pattern Assignment
NUC	Not under command

. . . continued

Acronym	Description
OOW	Officer of the Watch
OU	Ornstein-Uhlenbeck
PDAF	Probabilistic data association filter
PDF	Probability density function
PF	Particle filter
PHDF	Probability Hypothesis Density Filter
RIATM	Restricted, in ability to manoeuvre
RATDMA	Random Access Time Division Multiple Access
RF	Radio frequency
ROT	Rate of Turn
SaR	Search and Rescue
SAR	Synthetic Aperture Radar
SIAP	Single Integrated Air Picture
SOG	Speed over ground
SOLAS	Safety of Life at Sea
SOTMDA	Self-Organising Time Division Multiple Access
SQL	Structured Query Language
TRL	Technology Readiness Level
UKF	Unscented Kalman Filter
UN/LOCODE	United Nations Code for Trade and Transport Locations
UNECE	United Nations Economic Commission for Europe
UTC	Coordinated Universal Time
UTM	Universal Transverse Mercator
VHF	Very high frequency
VMS	Vessel Monitoring System
VTS	Vessel Traffic Services
WIG	Wing in Ground
WRS	World Registry of Ships

Appendix D

Gibbs sampling

To obtain a sample from the multivariate distribution,

$$\pi(\theta_1, \ldots, \theta_d), \tag{D.1}$$

$\pi(\boldsymbol{\theta})$ is the target distribution or joint posterior distribution.

The Gibbs sampler obtains a sample from $\pi(\boldsymbol{\theta})$ by successively and repeatedly simulating from the conditional distribution of each component of $\boldsymbol{\theta}$ given the other components $\pi(\theta_i, \theta_{-i})$.

D.1 The Gibbs Sampler Algorithm

1. Initialise with $\boldsymbol{\theta}_0$, with $\theta = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.
2. Simulate $\theta_1^{(1)}$ from the conditional distribution $\pi(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)})$.
3. Simulate $\theta_2^{(1)}$ from the conditional distribution $\pi(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)})$.
4. Continue...
5. Simulate $\theta_d^{(1)}$ from the conditional distribution $\pi(\theta_d | \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{d-1}^{(1)})$.
6. Iterate.

Under mild regularity conditions, convergence of the Markov chain to the stationary distribution $\pi(\boldsymbol{\theta})$ is guaranteed. Subsequent draws after a burn-in period (set of approx. 300 draws are discarded)

$$\pi(\boldsymbol{\theta}^{(1)}), \ldots, \pi(\boldsymbol{\theta}^{(J)}) \tag{D.2}$$

are realisations of the distribution $\pi(\boldsymbol{\theta})$.

Bayes theorem has the form

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \tag{D.3}$$

It is often impossible to solve the normalising integrals but with the Gibbs sampler, as long as it is possible to sample from each conditional posterior distribution, samples can be obtained from the joint posterior distribution without computing any integrals.

Once a sample $\boldsymbol{\theta}^{(i)}$ from $\pi(\theta|y)$ is acquired, any of features of the posterior distributions can be approximated using the empirical counterpart.

$$\mathbb{E}[t(\theta)|y] \approx \frac{1}{J}\sum_{j=1}^{J} t(\theta^{(j)}) \tag{D.4}$$

for any $t(\theta)$.

Appendix E

Additional Information

E.1 Taylor Series Expansion

A single variable $f(x)$ can be expanded around a given point x by the Taylor series

$$f(x+\delta x) = f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \frac{1}{3!}f'''(x)\delta x^3 + \dots$$

when δx is small, the higher order terms can be neglected and approximate $f(x+\delta x)$ to a quadratic function;

$$\begin{aligned} f(x+\delta x) &= f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \frac{1}{3!}f'''(x)\delta x^3 + \dots \\ &= f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \varepsilon(\delta x^3) \\ &\approx f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 \end{aligned} \tag{E.1}$$

or a linear function,

$$\begin{aligned} f(x+\delta x) &= f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \frac{1}{3!}f'''(x)\delta x^3 + \dots \\ &= f(x) + f'(x)\delta x + \varepsilon(\delta x^2) \\ &\approx f(x) + f'(x)\delta x \end{aligned} \tag{E.2}$$

The multivariate scalar function $f(x_1, \ldots, x_N) = f(\boldsymbol{x})$ can be expanded similarly by the Taylor series.

$$f(\boldsymbol{x} + \delta\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{N} \frac{\partial f(\boldsymbol{x})}{\partial x_i}\delta x_i + \frac{1}{2!}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\delta x_i \delta x_j + \ldots$$

which can be expressed in vector form;

$$f(\boldsymbol{X} + \delta\boldsymbol{X}) = f(\boldsymbol{X}) + g^T\delta\boldsymbol{X} + \frac{1}{2}\delta\boldsymbol{X}^T\boldsymbol{H}\delta\boldsymbol{X} + \ldots$$

where $\delta\boldsymbol{X} = [\delta x_1, \ldots, \delta x_N]^T$, g is the gradient function and $\boldsymbol{H}$ is the Hessian matrix.

$$g = g_f(\boldsymbol{X}) = \nabla f(\boldsymbol{X}) = \frac{d}{d\boldsymbol{X}} f(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\boldsymbol{X}) \\ \vdots \\ \frac{\partial}{\partial x_N} f(\boldsymbol{X}) \end{bmatrix}_{N\times 1}$$

$$\boldsymbol{H} = \boldsymbol{H}_f(\boldsymbol{X}) = \frac{d}{d\boldsymbol{X}} g_f(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(\boldsymbol{X}) & \ldots & \frac{\partial^2}{\partial x_1 \partial x_N} f(\boldsymbol{X}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_N \partial x_1} f(\boldsymbol{X}) & \ldots & \frac{\partial^2}{\partial x_N^2} f(\boldsymbol{X}) \end{bmatrix}_{N\times N}$$

The multivariable vector case can be adapted for the multivariable scalar case.

For a set of M multivariable functions $f_i(\boldsymbol{X})$ for $i = 1, \ldots, M$ expressed as

$$\boldsymbol{f}(\boldsymbol{X}) = [f_1(\boldsymbol{X}), \ldots, f_M(\boldsymbol{X})]^T$$

The Taylor series expansion for the i-th function component is

$$f_i(\boldsymbol{X} + \delta\boldsymbol{X}) = f_i(\boldsymbol{X}) + g_i^T\delta\boldsymbol{X} + \frac{1}{2}\delta\boldsymbol{X}^T\boldsymbol{H}_i\delta\boldsymbol{X} + \varepsilon(\|\delta\boldsymbol{X}\|^3)$$

The vector form can be written as

$$\boldsymbol{f}(\boldsymbol{X} + \delta\boldsymbol{X}) \approx \begin{bmatrix} f_1(\boldsymbol{X}) \\ \vdots \\ f_M(\boldsymbol{X}) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \ldots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1} & \ldots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \vdots \\ \delta x_N \end{bmatrix} = \boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{X})\delta\boldsymbol{X}$$

where $\boldsymbol{J_f}(\boldsymbol{X})$ is the Jacobian matrix defined on the function $\boldsymbol{f}(\boldsymbol{X})$.

$$\boldsymbol{J_f}(\boldsymbol{X}) = \frac{d}{d\boldsymbol{X}}\boldsymbol{f}(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_N} \end{bmatrix} \begin{bmatrix} f_1 & \dots & f_N \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix}_{M \times N}$$

The 2nd order term (not required for linear approximation but needed in covariance calculation) can no longer be expressed as a matrix as it is now a tensor.

The Hessian matrix for $f(\boldsymbol{X})$ can be calculated from the Jacobian of the gradient vector.

$$\boldsymbol{g}_f(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix}$$

$$\boldsymbol{J_g}(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix} = \boldsymbol{H_f}(\boldsymbol{X})$$

when $M > N$, the problem is over constrained.

Bibliography

[1] Joan Albiol, Jordi Robusté, Carles Casas, and Manel Poch. Biomass estimation in plant cell cultures using an extended Kalman filter. *Biotechnology progress*, 9(2):174–178, 1993.

[2] Tom Arnold, Mark J. Bertus, and Jonathan Godbey. A simplified approach to understanding the Kalman filter technique. *The Engineering Journal*, 53:140–155, 2008.

[3] Augusto Aubry, Paolo Braca, Enrica d'Afflisio, Antonio De Maio, Leonardo M. Millefiori, and Peter Willett. Optimal opponent stealth trajectory planning based on an efficient optimization technique. *IEEE Transactions on Signal Processing*, 69:270–283, 2020.

[4] Yaakov Bar-Shalom, Fred Daum, and Jim Huang. The probabilistic data association filter. *IEEE Control Systems Magazine*, 29(6):82–100, 2009.

[5] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[6] Yaakov Bar-Shalom and Xiao-Rong Li. *Multitarget-multisensor tracking: principles and techniques*, volume 19. YBs Storrs, CT, 1995.

[7] Yaakov Bar-Shalom and Edison Tse. Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5):451–460, 1975.

[8] Adam Bennett. HawkEye 360 begins first-of-its-kind commercial geolocation of radio frequency signals from space. Online: `https://www.prnewswire.com/news-`

releases/hawkeye-360-begins-first-of-its-kind-commercial-geolocation-of-radio-frequency-signals-from-space-300801703.html, accessed 13th April 2019.

[9] José M. Bernardo and Adrian F. M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[10] W Dale Blair. Multitarget tracking metrics for SIAP systems. In *Int. Conf. on Info Fusion-Fusion08*, 2008.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[12] Paolo Braca, Salvatore Maresca, Raffaele Grasso, Karna Bryan, and Jochen Horstmann. Maritime surveillance with multiple over-the-horizon HFSW radars: An overview of recent experimentation. *IEEE Aerospace and Electronic Systems Magazine*, 30(12):4–18, 2015.

[13] Laura M. Bradbury, Nathan G. Orr, Maria Short, Niels Roth, Arunas Macikunas, Balaji Kumar, Chris Short, Barbara Ham, and Robert E. Zee. exactview-9: Commissioning and on-orbit operation of a high performance AIS nanosatellite. In *14th International Conference on Space Operations*, 2016.

[14] Luigi Bruno, Paolo Braca, Jochen Horstmann, and Michele Vespe. Experimental evaluation of the range–doppler coupling on HF surface wave radars. *IEEE Geoscience and Remote Sensing Letters*, 10(4):850–854, 2012.

[15] Stephan Brusch, Susanne Lehner, Thomas Fritz, Matteo Soccorsi, Alexander Soloviev, and Bart van Schie. Ship surveillance with TerraSAR-X. *IEEE transactions on geoscience and remote sensing*, 49(3):1092–1103, 2010.

[16] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 12 1979.

[17] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[18] Samuele Capobianco, Leonardo M. Millefiori, Nicola Forti, Paolo Braca, and Peter Willett. Deep learning methods for vessel trajectory prediction based on recurrent neural networks. *arXiv preprint arXiv:2101.02486*, 2021.

[19] Gregory Castñnón and Lucas Finn. Multi-target tracklet stitching through network flows. In *2011 Aerospace Conference*, pages 1–7. IEEE, 2011.

[20] Neil Cater. Long range identification and tracking. *Journal of Ocean Technology*, 4(2):8–18, 2009.

[21] Centre of Excellence for Operations in Confined and Shallow Waters. Maritime situational awareness. Online: `https://www.coecsw.org/our-work/projects/maritime-situational-awareness/`, accessed 12th April 2019.

[22] Jie Chen and Arjun K. Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.

[23] Lingji Chen and Sarah E. Rumbley. Track stitching and approximate track association on a pairwise-likelihood graph. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVII*, volume 10646, page 1064602. International Society for Optics and Photonics, 2018.

[24] Y Cheng. Satellite-based ais and its comparison with LRIT. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 8(2), 2014.

[25] Kristina Chodorow. *MongoDB: the definitive guide: powerful and scalable data storage*. O'Reilly Media, Inc., 2013.

[26] Chee-Yee Chong. Graph approaches for data association. In *2012 15th International Conference on Information Fusion*, pages 1578–1585. IEEE, 2012.

[27] Pasquale Coscia, Paolo Braca, Leonardo M. Millefiori, Francesco A. N. Palmieri, and Peter Willett. Multiple Ornstein–Uhlenbeck processes for maritime traffic graph representation. *IEEE Transactions on Aerospace and Electronic Systems*, 54(5):2158–2170, 2018.

[28] CTI Digital. Royal Navy's NELSON. Online: `ctidigital.com/our-clients/nelson-royal-navy`, accessed 28th December 2019.

[29] Meghan Curran. Soft targets & black markets: Terrorist activities in the maritime domain. Technical report, One Earth Future, May 2019.

[30] Enrica d'Afflisio, Paolo Braca, and Peter Willett. Malicious ais spoofing and abnormal stealth deviations: A comprehensive statistical framework for maritime anomaly detection. *IEEE Transactions on Aerospace and Electronic Systems*, 2021.

[31] Clément Dechesne, Sébastien Lefèvre, Rodolphe Vadaine, Guillaume Hajduch, and Ronan Fablet. Ship identification and characterization in sentinel-1 SAR images with multi-task deep learning. *Remote Sensing*, 11(24):2997, 2019.

[32] Defense Mapping Agency. *Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems*, volume 8350. Defense Mapping Agency, 1987.

[33] Denbridge Marine. Denbridge Marine Limited. Online: `https://www.denbridgemarine.com/`, accessed 26th May 2016.

[34] Department for Transport. Shipping fleet statistics 2016. Technical report, UK Government, 2017.

[35] Nathan S. Dietrich, Robert A. Koyak, and Thomas W. Lucas. Algorithms. 2001.

[36] Jack Doyle. National Maritime Information Centre will monitor threat from sea. Online: `https://www.independent.co.uk/news/uk/crime/national-maritime-information-centre-will-monitor-threat-sea-1925293.html`, accessed 22nd October 2017.

[37] Exact Earth. exactAIS archive. Online: `https://www.exactearth.com/products/exactais-archive`, accessed 4th July 2018.

[38] Nicola Forti, Leonardo M. Millefiori, Paolo Braca, and Peter Willett. Random finite set tracking for anomaly detection in the presence of clutter. In *2020 IEEE Radar Conference (RadarConf20)*, pages 1–6. IEEE, 2020.

[39] W. G. Franke, R. Freyer, H. D. Gebauer, L. Oehme, and T. Schmitt. Enhancement of SPECT images using a 3d Kalman filter for evaluation of brain scintigrams. In *Journal of Nuclear Medicine*, volume 37, pages 947–947. SOC Nuclear Medicine Inc. 1850 Samuel Morse Dr., Reston, VA 22090-5316, 1996.

[40] Domenico Gaglione, Giovanni Soldi, Florian Meyer, Franz Hlawatsch, Paolo Braca, Alfonso Farina, and Moe Z. Win. Bayesian information fusion and multitarget tracking for maritime situational awareness. *IET Radar, Sonar & Navigation*, 14(12):1845–1857, 2020.

[41] Basil Germond. The geopolitical dimension of maritime security. *Marine Policy*, 54:137–142, 2015.

[42] Karl Granström, Antonio Natale, Paolo Braca, Giovanni Ludeno, and Francesco Serafino. Gamma gaussian inverse Wishart probability hypothesis density for extended target tracking using X-band marine radar data. *IEEE Transactions on Geoscience and Remote Sensing*, 53(12):6617–6631, 2015.

[43] GraphQL. GraphQL: A query language for your API. Online: `https://graphql.org/`, accessed 8th February 2019.

[44] Olaf Hartig and Jorge Pérez. An initial analysis of Facebook's GraphQL language. 2017.

[45] Mihály Héder. From NASA to EU: the evolution of the TRL scale in public sector innovation. *The Innovation Journal: The Public Sector Innovation Journal*, 22(2), 2017.

[46] David V. Hinkley. Inference about the change-point in a sequence of random variables. 1970.

[47] IHS Markit. Sea-web: The ultimate marine online database. Online: `https://ihsmarkit.com/products/sea-web-maritime-reference.html`, accessed 13th July 2019.

[48] IHS Markit. IHS Markit: Leading source of critical information. Online: `https://ihsmarkit.com/index.html`, accessed 19th January 2016.

[49] IHS Markit. AISLive: Maritime ship and vessel tracker. Online: `https://www.ihs.com/products/ais-live-ship-tracker.html`, accessed 20th December 2016.

[50] International Association of Marine Aids to Navigation and Lighthouse Authorities. Guideline no. 1082 on an overview of AIS, 2011.

[51] International Maritime Organisation. *International convention for the safety of life at sea*. International Maritime Organisation, 1974.

[52] International Maritime Organisation. IMO resolution A.600(15), 1987.

[53] International Maritime Organisation. *Annex to the International Convention for the Safety of Life at Sea*, chapter Chapter V: Safety of Navigation, pages 1–1. International Maritime Organisation, 2002.

[54] International Maritime Organisation. Guidelines for the onboard operational use of shipborne automatic identification systems (AIS), resolution A.917(22), 2002.

[55] International Maritime Organisation. Safety of navigation. *International Convention for the Safety of Life at Sea*, 2002.

[56] International Maritime Organisation. Maritime security and piracy. Online: `http://www.imo.org/en/OurWork/Security/Pages/MaritimeSecurity.aspx`, accessed 24th April 2016.

[57] International Maritime Organisation. Introduction to IMO. Online: `http://www.imo.org/en/About/Pages/Default.aspx`, accessed 31st January 2019.

[58] International Maritime Organization. Resolution MSC 298(87):establishment of distribution facility for the provision of LRIT information to security forces operating in waters in the Gulf of Aden and the Western Indian Ocean to aid their work in the repression of piracy and armed robbery against ships. Technical Report 87, International Maritime Organization, 2010.

[59] International Organization for Standardization. Glossary for ISO 3166.

[60] International Organization for Standardization. Road vehicles - electrical disturbances from conduction and coupling: Part 1: Definitions and general considerations. Technical report, International Organization for Standardization, Vernier, Geneva, Switzerland, 2015.

[61] International Telecommunication Union. M.1371-5:technical characteristics for an automatic identification system using time-division multiple access in the vhf maritime mobile frequency band, 2014.

[62] Clément Iphar, Aldo Napoli, and Cyril Ray. Detection of false ais messages for the improvement of maritime situational awareness. In *OCEANS 2015 - MTS/IEEE Washington*, pages 1–7, 2015.

[63] Oliver Louis Robert Jacobs. *Introduction to Control Theory*. Oxford University Press, Oxford ;, 2nd ed. edition, 1993.

[64] Thorben Janz, Andreas Leich, Marek Junghans, Kay Gimm, Shishan Yang, and Marcus Baum. Post-processing of multi-target trajectories for traffic safety analysis. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2018.

[65] Pierre Joly, Daniel Commenges, Catherine Helmer, and Luc Letenneur. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, 3(3):433–443, 2002.

[66] Simon J. Julier and Jeffrey K. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, University of Oxford, 1996.

[67] Simon J. Julier and Jeffrey K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. pages 182–193, 1997.

[68] Simon J. Julier, Jeffrey K. Uhlmann, and Hugh. F. Durrant-Whyte. A new approach for filtering nonlinear systems. In *Proceedings of 1995 American Control Conference - ACC'95*, volume 3, pages 1628–1632, 1995.

[69] Simon J. Julier, Jeffrey K. Uhlmann, and Hugh F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3):477–482, 2000.

[70] Rudolf E. Kálmán. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82(1):35–45, March 1960.

[71] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of change-points with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[72] Martin Kirscht and Carsten Rinke. 3D reconstruction of buildings and vegetation from synthetic aperture radar (SAR) images. In *MVA*, 1998.

[73] John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

[74] Richard Lane, Mark Briers, Thomas Cooper, and Simon Maskell. Efficient data structures for large scale tracking. *FUSION 2014 - 17th International Conference on Information Fusion*, 10 2014.

[75] Richard Lane, David Nevell, Steve Hayward, and Thomes W. Beaney. Maritime anomaly detection and threat assessment. *Proc. of the Int. Conf. on Information Fusion*, pages 1 – 8, 08 2010.

[76] Delphine Lautier. The Kalman filter in finance: An application to term structure models of commodity prices and a comparison between the simple and the extended filters. 01 2002.

[77] Eric Lefebvre and Christopher Helleur. Use of fuzzy logic for data fusion in a recognized maritime picture. In *Advances in Intelligent Systems, Fuzzy Systems, Evolutionary Computation (Proceedings of the 3rd WSES International Conference on Fuzzy Sets and Fuzzy Systems)*, pages 24–29, 2002.

[78] Tine Lefebvre, Herman Bruyninckx, and Joris De Schutter. Kalman filters for nonlinear systems: A comparison of performance. *International Journal of Control*, 77:639, 2004.

[79] Dimitrios Lekkas, Spyros Vosinakis, Charalambos Alifieris, and John Darzentas. Marinetraffic: Designing a collaborative interactive vessel traffic information system. Technical report, MarineTraffic, 2003.

[80] Larry J. Levy. The Kalman filter: Navigation's integration workhorse. *GPS World*, 8(9):65–71, September 1997.

[81] D. Y. Lin, E. J. Feuer, R. Etzioni, and Y. Wax. Estimating medical costs from incomplete follow-up data. *Biometrics*, 53(2):419–434, 1997.

[82] Yimeng Liu, Yizhi Wang, and Yi Jin. Research on the improvement of MongoDB auto-sharding in cloud environment. In *2012 7th international conference on Computer science & education (ICCSE)*, pages 851–854. IEEE, 2012.

[83] Lloyd's List Intelligence. Lloyd's list. Online: `https://lloydslist.maritimeintelligence.informa.com/`, accessed 18th June 2017.

[84] Ronald P.S. Mahler. Multitarget bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic systems*, 39(4):1152–1178, 2003.

[85] Larry M. Manevitz and Malik Yousef. One-class SVMs for document classification. *Journal of machine Learning research*, 2(Dec):139–154, 2001.

[86] Salvatore Maresca, Paolo Braca, Jochen Horstmann, and Raffaele Grasso. Maritime surveillance using multiple high-frequency surface-wave radars. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):5056–5071, 2013.

[87] Marine Traffic. Global ship tracking intelligence: Ais marine traffic. Online: `https://www.marinetraffic.com/`, accessed 1st June 2018.

[88] Marine Traffic. Global ship tracking intelligence: Ais marine traffic. Online: `https://www.marinetraffic.com/`, accessed 1st June 2018.

[89] Maritime and Coastguard Agency. MSN 1781 (M + F): The Merchant Shipping (Distress Signals and Prevention of Collisions) Regulations 1996. Technical Report 87, Maritime and Coastguard Agency, 2010.

[90] Rey Q. Masinsin. The single integrated air picture: Building synergy for theater air and missile defense? Technical report, Marine Corps Command and Staff Coll., Quantico, VA, 2000.

[91] Simon Maskell. Track analytics for effective triage of wide area surveillance data: Phase 2a final report. Technical report, University of Liverpool, 2020.

[92] Mike Mathis, Harry Dutchyshyn, and Jeffery W. Wilson. Single integrated air picture (SIAP) progress, plans, and recommendations. Technical report, Single Integrated Air Picture System Engineering Task Force, Arlington, VA, 2002.

[93] Peter S. Maybeck. *Stochastics Models, Estimation, and Control: Introduction*, volume 1. Academic Press, 1979.

[94] Efim Mazor, Amir Averbuch, Yaakov Bar-Shalom, and Joshua Dayan. Interacting multiple model methods in target tracking: A survey. *Aerospace and Electronic Systems, IEEE Transactions on*, 34(1):103 – 123, 02 1998.

[95] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.

[96] Andrew Metrick and Kathleen H. Hicks. Contested seas: Maritime domain awareness in northern Europe. Technical report, Center for Strategic & International Studies, 2018.

[97] Ryszard K. Miler and Andrzej Bujak. exactEarthSatellite – ais as one of the most advanced shipping monitoring systems. In Jerzy Mikulski, editor, *Activities of Transport Telematics*, pages 330–337, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[98] Leonardo M. Millefiori, Paolo Braca, Karna Bryan, and Peter Willett. Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction. *IEEE Transactions on Aerospace and Electronic Systems*, 52(5):2313–2330, 2016.

[99] Leonardo M. Millefiori, Paolo Braca, and Peter Willett. Consistent estimation of randomly sampled Ornstein–Uhlenbeck process long-run mean for long-term target state prediction. *IEEE Signal Processing Letters*, 23(11):1562–1566, 2016.

[100] Ministry of Defence, Department for Transport, Foreign & Commonwealth Office, and Home Office. National strategy for maritime security. Technical report, UK Government, 2014.

[101] Sharad Nagappa, Daniel E. Clark, and Ronald Mahler. Incorporating track uncertainty into the OSPA metric. In *14th International Conference on Information Fusion*, pages 1–8. IEEE, 2011.

[102] National Maritime Information Centre. National Maritime Information Centre. Online: `http://www.nmic.org.uk/`, accessed 6th March 2018.

[103] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2-3):103–134, 2000.

[104] Lisa Otto. *Global Challenges in Maritime Security: An Introduction*. Springer Nature, 2020.

[105] Ewan S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[106] Athanasios Papoulis and H. Saunders. Probability, random variables and stochastic processes. 1989.

[107] Charles P. Quesenberry, Jr, Bruce Fireman, Robert A. Hiatt, and Joseph V. Selby. A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome. *American Journal of Public Health*, 79(12):1643–1647, 1989.

[108] Abu Sajana Rahmathullah, Ángel F. García-Fernández, and Lennart Svensson. Generalized optimal sub-pattern assignment metric. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8. IEEE, 2017.

[109] M. P. Rajan and Jimson Mathew. Kalman filter and financial time series analysis. In Jimson Mathew, Priyadarshan Patra, Dhiraj K. Pradhan, and A. J. Kuttyamma, editors, *Eco-friendly Computing and Communication Systems*, pages 339–351, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[110] Herbert E. Rauch, F. Tung, and Charlotte T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

[111] Alberto Regattieri, R. Manzini, and Daria Battini. Estimating reliability characteristics in the presence of censored data: A case study in a light commercial vehicle manufacturing system. *Reliability Engineering & System Safety*, 95(10):1093–1102, 2010.

[112] Marco Reggiannini, Marco Righi, Marco Tampucci, Luigi Bedini, Claudio Di Paola, Massimo Martinelli, Costanzo Mercurio, and Emanuele Salerno. Remote sensing for maritime monitoring and vessel prompt identification. In Kazimierz Choroś, Marek Kopel, Elżbieta Kukla, and Andrzej Siemiński, editors, *Multimedia and Network Information Systems*, pages 343–352, Cham, 2019. Springer International Publishing.

[113] Maria Riveiro and Göran Falkman. Supporting the analytical reasoning process in maritime anomaly detection: Evaluation and experimental design. In , pages 170–178, 07 2010.

[114] Maria Riveiro, Göran Falkman, and Tom Ziemke. Improving maritime anomaly detection and situation awareness through interactive visualization. In *Proceedings of the 11th International Conference on Information Fusion, FUSION 2008*, pages 1 – 8, 01 2008.

[115] Sheldon M. Ross. *Introductory statistics*. Academic Press, 2017.

[116] Hanan Samet. The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)*, 16(2):187–260, 1984.

[117] Paul Scerri, Robin Glinton, Sean Owens, David Scerri, and Katia Sycara. Geolocation of RF emitters by many UAVs. *American Institute of Aeronautics and Astronautics*, 2007.

[118] Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE transactions on signal processing*, 56(8):3447–3457, 2008.

[119] Andrew John Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512, 1974.

[120] Ritesh Singh and Keshab Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, 2(4):145, 2011.

[121] James A. Slater and Stephen Malys. WGS 84 - Past, Present and Future. In Fritz K. Brunner, editor, *Advances in Positioning and Reference Frames*, pages 1–7, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[122] John P. Snyder. Map projections: A working manual. Technical Report 1395, U.S. Geological Survey, 1987.

[123] SOLAS. International convention for the safety of life at sea (SOLAS), 1974.

[124] Harold W. Sorenson. Least-squares estimation: from Gauss to Kalman. *IEEE Spectrum*, 7(7):63–68, 1970.

[125] Mengwei Sun, David Cormack, and James Hopgood. Tracking and sensing. In *UDRC-EURASIP Summer School Presentations 2019*, 2019.

[126] Ruben Taelman, Miel Vander Sande, and Ruben Verborgh. GraphQL-LD: linked data querying with GraphQL. In *ISWC2018, the 17th International Semantic Web Conference*, pages 1–4, 2018.

[127] Bruno O.S. Teixeira, Jaganath Chandrasekar, Leonardo A.B. Tˆôrres, Luis A. Aguirre, and Dennis S. Bernstein. State estimation for linear and non-linear equality-constrained systems. *International Journal of Control*, 82(5):918–936, 2009.

[128] Paul A. Thomas, Jordi Barr, Bhashyam Balaji, and Kruger White. An open source framework for tracking and state estimation ('stone soup'). In Ivan Kadar, editor, *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVI*, volume 10200, pages 62 – 71. International Society for Optics and Photonics, SPIE, 2017.

[129] Paul A. Thomas, Jordi Barr, Steven Hiscocks, Charlie England, Simon Maskell, Bhashyam Balaji, and Jason Williams. Stone soup: An open-source framework for tracking and state estimation. In *ISIF Perspectives Magazine*, volume 2, pages 14–19, 2019.

[130] Kerry Trentelman, Rebecca Rafferty, Adam Saulwick, and Aaron Ceglar. Information fusion for maritime domain awareness: Illegal fishing detection (poster). In *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 134–139, 2019.

[131] George Eugene Uhlenbeck and Leonard Salomon Ornstein. On the theory of the Brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.

[132] UK Government. The UK national strategy for maritime security. Online: `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/310987/20140508_NSMS.pdf`, May 2014.

[133] UK Government. National strategy for maritime security. Online: `https://www.gov.uk/government/publications/national-strategy-for-maritime-security`, accessed 24th March 2017.

[134] United Nations Economic Commission for Europe. UN/LOCODE. Online: `http://www.unece.org/cefact/locode/welcome.html`, accessed 27th January 2020.

[135] U.S. Geological Survey. The Universal Transverse Mercator (UTM) Grid. Fact Sheet 077-01, U.S. Geological Survey, 2001.

[136] Lodewyk J. Van der Merwe and Johan Pieter de Villiers. Track-stitching using graphical models and message passing. In *Proceedings of the 16th International Conference on Information Fusion*, pages 758–765. IEEE, 2013.

[137] Vessel Finder. Free AIS ship tracking of marine traffic. Online: `https://www.vesselfinder.com/`, accessed 1st May 2018.

[138] Gemine Vivone, Leonardo M. Millefiori, Paolo Braca, and Peter Willett. Performance assessment of vessel dynamic models for long-term prediction using heterogeneous data. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6533–6546, 2017.

[139] Paul Votruba, Rich Nisley, Ron Rothrock, and Brett Zombro. Single integrated air picture (SIAP) metrics implementation. Technical report, Single Integrated Air Picture System Engineering Task Force, Arlington, VA, 2001.

[140] Daphna Weinshall, Gal Levi, and Dmitri Hanukaev. LDA topic model with soft assignment of descriptors to words. *International Conference on Machine Learning*, pages 711–719, 2013.

[141] Curt Wells. *The Kalman filter in finance*, volume 32. Springer Science & Business Media, 2013.

[142] Philip Whittaker, Martin Cohen, David Hall, and Luis Gomes. An affordable small satellite SAR mission. In *Proceedings of the 8th IAA Symposium on Small Satellites for Earth Observation*, 2011.

[143] Harindra C. Wijeysundera, Xuesong Wang, George Tomlinson, Dennis T. Ko, and Murray D. Krahn. Techniques for estimating health care costs with censored data: an overview for the health services researcher. *ClinicoEconomics and outcomes research: CEOR*, 4:145, 2012.

[144] James Wright. Disambiguating cooperative sensor data. In *Cranfield University (2018): 2018 Defence and Security Doctoral Symposium (DSDS19) in conjunction with DSTL, AWE, Department for Transport and NCSC: Symposium outputs*, Poster. Cranfield University, 2018.

[145] Jianhua Yin and Jianyong Wang. A Dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, 2014.

[146] Linxiang Zhang. Data and content analysis for social network using LDA text model. *Journal of Physics: Conference Series*, 1213:022035, 06 2019.

[147] Hao Zheng and Hongwei Wu. A novel LDA and PCA-based hierarchical scheme for metagenomic fragment binning. In *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 53 – 59, 05 2009.

[148] Yifan Zhou, James Wright, and Simon Maskell. A generic anomaly detection approach applied to mixture-of-unigrams and maritime surveillance data. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–6, 2019.